Cis-regulation and genetic control of gene expression in neuroblastoma

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

im Fach Biologie

eingereicht an der

Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

Christian Martin Burkert (M. Sc.)

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät:
Prof. Dr. Dr. Christian Ulrichs

Gutachter/innen:
1. Prof. Dr. Uwe Ohler
2. Prof. Dr. Nils Blüthgen
3. Prof. Dr. Niko Beerenwinkel

Tag der mündlichen Prüfung: 18.06.2021

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected.
I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Contributions

The results presented in this thesis are based on contributions of collaborators. Unless otherwise stated Christian Martin Burkert conducted all work under supervision of Dr. Roland F. Schwarz (Max Delbrück Center for Molecular Medicine [MDC], Berlin, Germany) and Prof. Dr. Uwe Ohler (MDC, Berlin, Germany). Collection of samples and clinical data was performed by the Terminate NB consortium headed by Prof. Dr. Angelika Eggert (Charité – Universitätsmedizin Berlin, Germany). Prof. Dr. Johannes Schulte (Charité – Universitätsmedizin Berlin, Germany) provided sequencing data and helped with the interpretation of clinical annotations. Samples jointly analyzed with those from the Terminate NB consortium were obtained from the publication Peifer et al. 2015. Alignments of normal WGS, tumor WGS, tumor RNA-seq and somatic single nucleotide variant calls were created by the Core Unit Bioinformatics (CUBI) of the Berlin Institute of Health (Berlin, Germany) under supervision of Dr. Dieter Beule. Dr. Jörn Tödling (AG Schulte, Charité Universitätsmedizin Berlin, Germany) provided somatic structural variant calls of neuroblastoma primary tumors that are analyzed in chapter 3. Richard P. Koche, PhD (Memorial Sloan Kettering Cancer Center, New York, USA) and Dr. Anton Henssen (Charité – Universitätsmedizin Berlin, Germany) provided alignments of Circle-seq reads and ecDNA regions per sample analyzed in chapter 4. Remo Monti (AG Ohler, MDC, Berlin, Germany) integrated the fastlmm and PEER methods into the data processing pipeline. These methods were used to produce data that is analyzed in chapter 3 and 5. Christiane Weber performed an initial analysis on a subset of samples that provided valuable input for the survival analysis of copy-number alterations presented in chapter 3. Dr. Dubravka Vucicevic (AG Ohler, MDC, Berlin, Germany) prepared ATAC-seq libraries and Dr. Scott Lacadie (AG Ohler, MDC, Berlin, Germany) generated corresponding peak calls and signal tracks that are analyzed in chapter 5.

In this work the singular first person (“I”, “my”) refers to the author’s work and his perspective. Deviating from this, the plural (“we”, “our”) is consistently used throughout methods, results and discussions of chapters 3, 4 and 5 to acknowledge contributions. Despite this, most of the work and perspectives referred to in these chapters are still those of the author. Detailed contributions are listed at the beginning of chapters where applicable.

Parts of this thesis have been published as research articles, other parts are included in a manuscript in preparation. Chapter 4 includes results published in Koche et al. 2020.

Acknowledgements

I would like to thank Dr. Roland F. Schwarz and Prof. Dr. Uwe Ohler for project supervision and funding, and my thesis advisory committee members Prof. Dr. Nils Blüthgen and Prof. I would like to thank Dr. Sebastian Waszak, Christiane Weber, Dr. Anton Henssen, Dr. Florian Massip, Stella Debiase and all members of the Schwarz and Ohler laboratories for many useful discussions. I thank Dr. Johannes Schulte for continuous feedback on my research project. I thank Dr. Eric Blanc, Nina Thiessen, Dr. Manuel Holtgrewe and all members of the CUBI team at the Berlin Institute for Health, who helped with access to the data and the distributed compute environment and who patiently answered all of my questions regarding these resources. I thank Michaela Kolbe and Dr. Michaela Herzig for the administrative support at MDC, as well as Dr. Cordelia Arndt-Sullivan and Jana Lahmer for the administrative support at Humboldt Universität zu Berlin. I thank Ursula Burkert for proofreading this thesis. Finally, I would like to thank Dr. Marlene Thielecke, my entire family and my friends for their personal support.

Zusammenfassung

Genregulation kontrolliert Phänotypen im Kontext von Gesundheit und Krankheit. In Krebszellen moduliert das Zusammenspiel zwischen Keimbahnvariation, genetischen Aberrationen und epigenetischen Faktoren die Genexpression in cis. Das Neuroblastom ist eine Krebserkrankung, die häufig im Kindesalter auftritt und aus entarteten Vorläuferzellen des sympathischen Nervensystems entsteht. Es ist gekennzeichnet durch eine geringe Anzahl rekurrenter exonischer Mutationen, aber häufiger Veränderungen der somatischen Kopienzahl, einschließlich Genamplifikationen auf extrachromosomaler zirkulärer DNA. Bisher ist wenig darüber bekannt, wie lokale genetische und epigenetische Faktoren Gene im Neuroblastom regulieren und krankheitsspezifische Phänotypen verursachen. In dieser Arbeit kombiniere ich die allelspezifische Analyse von Sequenzierungsdaten ganzer Genome (WGS), Transkriptome und zirkulärer DNA von Neuroblastom-Patienten, um genetische und cis-regulatorische Effekte zu charakterisieren und Keimbahnregulationsvarianten durch cis-QTLs-Kartierung und Chromatinprofile zu priorisieren. Außerdem zeige ich, dass Dosis-Effekte, die durch somatische Kopienzahl verursacht werden, andere lokale genetische Effekte dominieren und wichtige Signalwege regulieren. Diese sind unter anderem an der Aufrechterhaltung der Telomere, der genomischen Stabilität und an neuronalen Prozessen beteiligt. Genamplifikationen zeigen starke Dosis-Effekte und befinden sich häufig auf großen und nicht auf kleinen extrachromosomalen zirkulären DNAs. Die Analyse zeigt, dass der Verlust von 11q zu einer Hochregulation von H3.3 und H2A Histonvarianten durch die Gene H3F3B und H2AFJ in Tumoren mit alternativer Verlängerung der Telomere (ALT) führt, und dass kooperative Effekte somatischer Strukturvarianten und erhöhter somatischer Kopienzahl das TERT Gen hochregulieren. Weitere Erkenntnisse sind, dass 17p-Ungleichgewichte der Kopienzahl und die damit verbundene Herunterregulierung einer neuronalen Gensignatur sowie die Hochregulierung des genomisch geprägten Gens RTL1 durch Kopienzahl-unabhängige allelische Dosis-Effekte mit einer ungünstigen Prognose verbunden sind. Die cis-QTL-Analyse bestätigt eine zuvor beschriebene Regulation des LMO1 Gens durch einen Super-Enhancer-Risikopolymorphismus und charakterisiert das regulatorische Potenzial weiterer GWAS-Risiko-Loci. Diese Arbeit unterstreicht die Bedeutung von Dosis-Effekten im Neuroblastom und liefert eine detaillierte Übersicht regulatorischer Varianten, die in dieser Krankheit aktiv sind.

Abstract

Gene regulation controls phenotypes in health and disease. In cancer, the interplay between germline variation, genetic aberrations and epigenetic factors modulate gene expression in cis. The childhood cancer neuroblastoma originates from progenitor cells of the sympathetic nervous system. It is characterized by a sparsity of recurrent exonic mutations but frequent somatic copy-number alterations, including gene amplifications on extrachromosomal circular DNA. So far, little is known on how local genetic and epigenetic factors regulate genes in neuroblastoma to establish disease phenotypes. I here combine allele-specific analysis of whole genomes, transcriptomes and circular DNA from neuroblastoma patients to characterize genetic and cis-regulatory effects, and prioritize germline regulatory variants by cis-QTLs mapping and chromatin profiles. The results show that somatic copy-number dosage dominates local genetic effects and regulates pathways involved in telomere maintenance, genomic stability and neuronal processes. Gene amplifications show strong dosage effects and are frequently located on large but not small extrachromosomal circular DNAs. My analysis implicates 11q loss in the upregulation of histone H3.3 and H2A variant genes H3F3B and H2AFJ in tumors with alternative lengthening of telomeres and cooperative effects of rearrangements and somatic copy-number gains in the upregulation of TERT. Both 17p copy-number imbalances and associated downregulation of neuronal genes as well as upregulation of the imprinted gene RTL1 by copy-number-independent allelic dosage effects is associated with an unfavorable prognosis. cis-QTL analysis confirms the previously reported regulation of the LMO1 gene by a super-enhancer risk polymorphism and characterizes the regulatory potential of additional GWAS risk loci. My work highlights the importance of dosage effects in neuroblastoma and provides a detailed map of regulatory variation active in this disease.

Table of contents

List of figures

List of tables

Abbreviations

AEI - Allelic expression imbalance
ALT - Alternative lengthening of telomeres
ANOVA - Analysis of variance
AR gene - Allelic regulated gene
ASCN - Allele-specific copy-number
ASCN - Allele-specific copy-number
ASE - Allele-specific expression
aseQTL - ASE quantitative trait locus
ATAC-seq - Assay for transposase-accessible chromatin with sequencing
BAF - B-allele frequency
bp - Basepairs
cDNA - Complementary DNA
CGH - Comparative genomic hybridization
ChIP-seq - Chromatin immunoprecipitation sequencing
CI - Confidence interval
CIMP - CpG island methylator phenotype
CIN - Chromosomal instability
Circle-seq - eccDNA sequencing
CN - Copy-number
CNV - Copy-number variation
CRE - Cis-regulatory element
CRISPR - Clustered regularly interspaced short palindromic repeats
DHS - DNase I-hypersensitive site
DNA - Deoxyribonucleic acid
DNAse-seq - DNase I hypersensitive sites sequencing
eccDNA - (see ecDNA)
ecDNA - Extrachromosomal circular DNA
EMT - Epithelial to mesenchymal transition
eQTL - Expression quantitative trait locus
FDR - False discovery rate
FISH - Fluorescence in situ hybridization
FWER - Family-wise error rate
GIN - Genomic instability
GO - Gene ontology
GWA - Genome-wide association
GWAS - Genome-wide association study
HWP - Hardy-Weinberg principle
hetSNP - heterozygous SNP
ICGC - International cancer genome consortium
ICR - Imprinting control region
INSS - International neuroblastoma staging system
ITH - Intra-tumor heterogeneity
IQR - Inter quartile range

LogR - Log ratio
LOH - Loss of heterozygosity
MAF - Minor allele frequency
MMEJ - Microhomology-mediated end joining
mRNA - Messenger-RNA
NB - Neuroblastoma
NHEJ - Non-homologous end joining
NGS - Next-generation sequencing
NMD - Nonsense-mediated (mRNA-)decay
QTL - Quantitative trait locus
RNA - Ribonucleic acid
RNA-seq - RNA sequencing
rRNA - Ribosomal RNA
SCNA - Somatic copy-number alteration
SD - Standard deviation
shRNA - Small hairpin RNA
SNP - Single nucleotide polymorphism
SNV - Single nucleotide variant
SV - Structural variation
TAD - Topologically associating domain
TERTr - Telomerase reverse transcriptase rearrangement
TCGA - The cancer genome atlas
TF - Transcription factor
TFBS - Transcription factor binding site
TSS - Transcription start site
UTR - Untranslated region
WGS - Whole genome sequencing

1 Introduction

The first draft of the human genome was published 18 years ago. This landmark development opened up the possibility for a variety of applications in the field of genomics. The uprise of high throughput technologies, such as next-generation sequencing (NGS), in the past decades, made it possible to efficiently characterize and quantify genetic information at a genome-wide scale. With help of these technologies and the human genome sequence, we can today address research questions that were inconceivable just a decade ago. The ongoing research effort in genomics and related disciplines helps us to understand the molecular components, processes, and elements that are employed by cells to use and maintain the genetic information in our DNA. The way this cellular information is processed is variable in development and between cells of different tissues, despite that all cells of an individual possess the same genome. Transcriptomic studies examine to what extent DNA is transcribed to RNA, a process that is essential for the cell to express its genetic code, for example by translating genes into proteins. Studies of the transcriptome (the entirety of RNAs) revealed that genes are expressed differently between environments, cell types, and in development. Around 20,000 genes in the human genome encode for proteins. However, these protein-coding genes account for only a small proportion of the genetic code. The vast majority of the genome is “non-coding”, but these regions have profound consequences on how genes are expressed. Non-coding genetic elements were identified that underlie the regulation of genes in cis. Genes and cis-regulatory elements undergo epigenetic modifications, which do not alter their DNA sequence but can influence their regulatory potential. It became clear that to gain an understanding of how genes are regulated we must not only study DNA sequence but also consider the state and structure of chromatin, the complex of protein, and DNA in the nucleus. The field of epigenomics investigates how chromatin controls gene expression and constitutes a molecular environment that allows genes to be regulated differently throughout development and between cell types. We have only just begun to understand how the cell’s interpretation of the genome translates into cellular phenotypes and subsequently into complex traits.

High throughput genomics also allows us to characterize and compare individual genomes and investigate the genetic basis of health and disease. It has been used to characterize heritable genetic variation in the germline of individuals, such as single nucleotide polymorphisms (SNPs) but also larger structural variants. Investigation of sequence

differences between individuals disease-associated variants, which are frequently located in the non-coding genome and are expected to alter the regulation of genes. Cancer is sometimes referred to as a disease of the genome because the DNA of malignant cells harbors a multitude of somatic variants that were found to promote the disease. Somatic variants are not inherited but acquired in somatic tissues after fertilization. Somatic variants include single nucleotide variants (SNVs), small insertions and deletions, somatic structural variants (SVs), and larger somatic copy-number alterations (SCNAs). In healthy cells, genetic information lies on 22 chromosomal pairs and two sex chromosomes (the karyotype). Somatic alterations in cancer can produce abnormal karyotypes, with an altered number of chromosomal copies or fusions between different chromosomes. Cancer cells can also contain genetic material that is separate from chromosomes, such as extrachromosomal circular DNAs (ecDNAs), molecules that are able to transfer gene copies to daughter cells independent from mitotic chromosomal segregation. Comparisons of the tumor and normal genomes identified cancer-associated somatic variation both in coding and non-coding regions. While somatic variants in protein-coding genes can convey their effect by altering protein structure, larger copy-number (CN) and non-coding variants are believed to be involved in the deregulation of gene expression.

In this thesis, I will investigate the genetic and cis-regulatory consequences of germline and somatic variations in the childhood cancer neuroblastoma. By integrating genomic, transcriptomic, and epigenomic data from NGS, we will gain insights into disease mechanisms and their association with malignant phenotypes. I hope that this work will contribute to the challenging mission of finding better treatments and hopefully one day a cure for this deadly disease.

Outline

I here gave a short introduction to the thesis. In the following chapter 2 I will describe the biological and technical background of my studies. First, I will introduce the disease neuroblastoma as well as concepts of gene regulation with a focus on genetic and cis-regulatory mechanisms. I will then emphasize aspects of gene regulation that are of particular interest in cancer and summarize previous findings on gene regulation in neuroblastoma. The regulatory potential of extrachromosomal circular DNAs and the control of telomere maintenance mechanisms are of particular interest in the characterization of neuroblastoma tumors and I will highlight these topics in two dedicated sections. I will also

describe the fundamental technical principles that underlie the analysis methods and finally summarize my research objectives.

Results are described in three separate chapters, each focused on a particular aspect of my analysis. In chapter 3 I will first report genetic variants in neuroblastoma tumors and their donors, quantify the impact of different classes of variants on global gene expression variability and specifically relate regulatory effects of somatic copy-number variation to telomere maintenance and survival. To shed light on the regulatory role of extrachromosomal circular DNA I will analyze these structures in a subset of tumors in chapter 4. To that end, I will investigate the interplay between circular DNA, somatic copy-number and allelic regulation. In chapter 5 is focused on germline regulatory aspects. To this end I will first establish a panel of candidate cis-regulatory variants that is active in neuroblastoma tumors and priotitize functional variants by epigenetic observations in cell lines. I will then examine the results at previously reported neuroblastoma susceptibility loci in order to determine potential regulatory mechanisms in disease predisposition.

In each chapter I summarize and discuss the results in the context of previous reports on neuroblastoma biology and gene regulation in this cancer. Finally, in chapter 6 I will conclude the discussion and suggest future research directions.

2 Background

This chapter provides the biological and technical background of the work presented. I will introduce the disease neuroblastoma, general principles of gene regulation, such as regulatory elements and chromatin state, and present previous findings on gene regulation in this disease. I will discuss ecDNA, genetic elements which are of particular interest in neuroblastoma, as they are involved in regulation of the MYCN oncogene through copy-number amplifications. Telomere maintenance is a requirement for replicative immortality in cancer and this work will later highlight regulatory aspects of two distinct telomere maintenance mechanisms, which will also be introduced here. A section focused on methodology will provide the technical background for the methods applied later in the analyses. It comprises a description of the DNA sequencing technology that was used to generate the data, approaches to allele-specific analysis of DNA and RNA, and basic principles of association testing between germline SNPs and quantitative traits. Finally, I will define research objectives that will be addressed by the analyses presented and discussed in subsequent chapters.

2.1 Neuroblastoma

Neuroblastoma (NB) is a tumor of the sympathetic nervous system in early childhood. It is the most common extracranial solid tumor in children accounting for 6-10% of cancer diagnoses (Stiller and Parkin 1992) and 9% of pediatric cancer deaths (Smith et al. 2010). Incidence is highest in the first year of life (25-50 cases per million), where 30% of diagnoses are made, and declines with age, with only 5% of cases diagnosed in patients older than ten years (Stiller and Parkin 1992). Age at diagnosis is associated with higher risk. Patients early diagnosed with Neuroblastoma have a good prognosis, but later diagnoses are associated with worse outcomes. Five year survival rate of patients diagnosed before their first birthday is 88% compared to 65% for diagnoses made in children between ages 1–14 (Smith et al. 2010).

The primary site of the neoplasm is most frequently the adrenal gland (47%). Still, primary tumors may also be located (with decreasing frequencies) in the retroperitoneal abdomen, thorax, pelvis, and neck (Vo et al. 2014). Embryonal tumors, like neuroblastoma, originate from undifferentiated cells during organ development. Neuroblastoma arises from neural

crest cells, a progenitor cell type of the sympathetic nervous system. During embryonic development, neural crest cells undergo an epithelial to mesenchymal transition (EMT) and become a population of motile cells. These cells migrate from the neural tube to other parts of the embryo, differentiating into a variety of cell types, including cells of sympathetic ganglia and the adrenal gland (Bronner and Simões-Costa 2016; Matthay et al. 2016), organs in which neuroblastoma tumors are frequently found.

Neuroblastoma is a disease with strong clinical heterogeneity reflected in contrasting survival probabilities between phenotypes. Tumor staging systems classify phenotypes to predict outcome. Classification into stages 1, 2A, 2B, 3, 4 and 4S according to the international neuroblastoma staging system (INSS) is performed on the basis of tumor confinement to the primary site, the involvement of local lymph nodes and the dissemination of tumor to distant body parts (Brodeur et al. 1988, 1993). Stages 1-3 characterize localized or unilateral tumors with and without the involvement of local lymph nodes. Stages 4 and 4S indicate the dissemination of the disease to distant body parts. Stage 4S is reserved for tumors of children younger than one year of age with dissemination limited to skin, liver, and/or with minimal bone marrow infiltration. Patients with tumors stage 1-3 and 4S have a favorable prognosis with five-year survival probabilities of 73% or higher. In contrast, stage 4 tumors are associated with worse outcomes and a five year survival probability as low as 33% (V. Castel et al. 1999). Different treatment regimes reflect the clinical heterogeneity of the disease. While stage 4 tumors are often treated with intensive radio- and chemotherapy after resection, some tumors regress spontaneously, a phenomenon commonly seen in stage 4S tumors (Evans et al. 1980; J. Pritchard and Hickman 1994). Indeed, stage 4S tumors were associated with excellent survival rates independent of surgical intervention (Katzenstein et al. 1998).

Alongside tumor staging, genetic biomarkers became important criteria for risk stratification. Most importantly, DNA amplification of the transcription factor MYCN was associated with poor prognosis (Brodeur et al. 1984; Seeger et al. 1985). MYCN is amplified in approximately 20% of cases (Matthay et al. 2016; Kaczówka et al. 2018) and is assessed in routine clinical diagnostic by fluorescence in situ hybridization (FISH). Amplification of the MYCN oncogene was found to be associated with worse survival rates in lower stage tumors (Wilson et al. 1991; Bagatell et al. 2009), underlining the importance of genetic biomarkers in risk stratification. Together with MYCN amplification, larger chromosomal features are known to be common somatic aberrations in neuroblastoma. Frequent losses of chromosome arms

1p and 11q (Brodeur, Sekhon, and Goldstein 1977; Brodeur et al. 1981; J. M. Maris et al. 1995, 2001) as well as 17q gains (Gilbert et al. 1984; Meddeb et al. 1996) were found in neuroblastoma tumors and cell lines. Losses of chromosome arm 1p and 11q were linked to decreased overall survival (J. M. Maris et al. 1995; Janoueix-Lerosey et al. 2009). Figure 1 shows the risk stratification scheme of the German neuroblastoma trial NB2004 (Oberthuer et al. 2009).

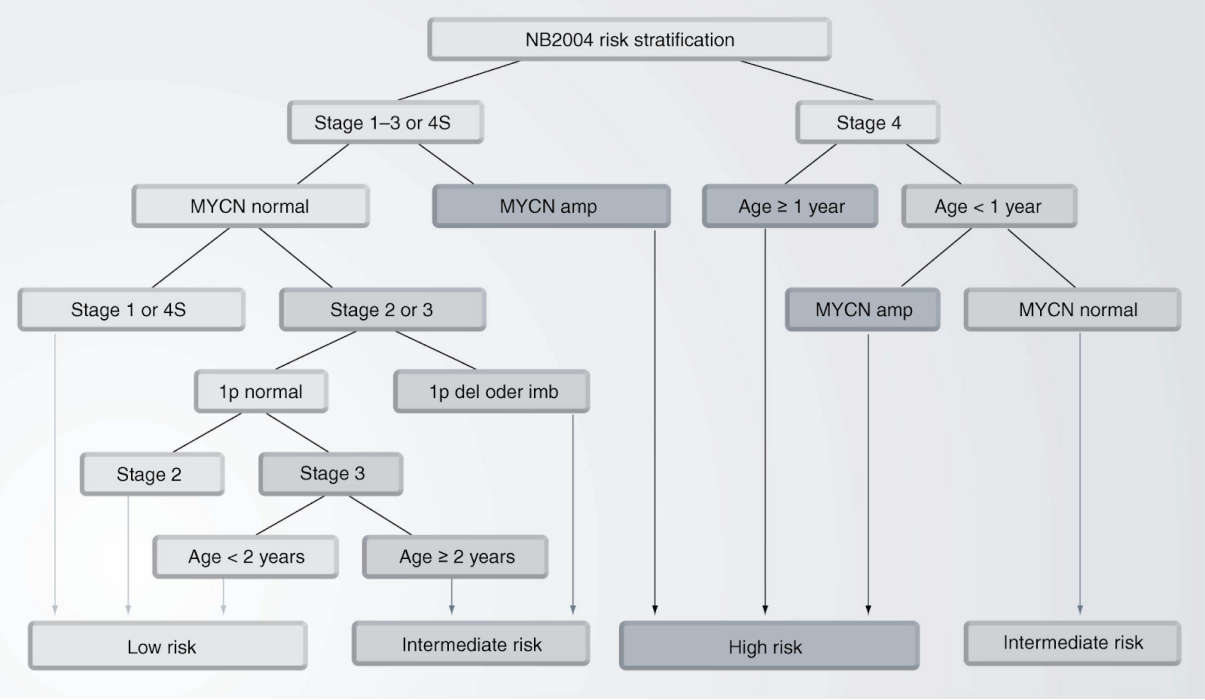


Figure 1: Risk stratification scheme from the neuroblastoma trial NB2004. Republished with permission of Future Medicine Ltd., from 'Molecular characterization and classification of neuroblastoma', André Oberthuer, Jessica Theissen, Frank Westermann, Barbara Hero & Matthias Fischer; Future Oncology Vol. 5, No. 5 (2009); permission conveyed through Copyright Clearance Center, Inc.

Sequencing studies found rare germline- and frequent somatic gain-of-function mutations or amplification of the receptor tyrosine kinase ALK (Mossé et al. 2008; R. E. George et al. 2008). While somatic mutations in ALK are relatively common (12% of high risk tumors), few other genes were detected to be recurrently affected by somatic mutations (R. E. George et al. 2008). However, a pathway analysis conducted by George et al. indicated an enrichment of genes involved in the RAS-MAPK signaling cascade, suggesting that functional mutations converge on this pathway. Clinical assessment of ALK and RAS-MAPK mutations may in the future open up possibilities for treatment by precision medicine approaches for a subset of patients.

Whole genome sequencing (WGS) studies have identified additional somatic aberrations. An array of exons of the chromatin remodeler ATRX was found to be frequently deleted in metastatic neuroblastoma (Cheung et al. 2012). Ceung and colleagues found the deletion to be associated with the absence of ATRX protein in the nucleus and long telomeres. Telomerase is a ribonucleoprotein involved in telomere lengthening, that is active during embryonic development but inactive in most somatic tissues after birth (W. E. Wright et al. 1996). Telomerase activity was found to be increased in neuroblastoma compared to normal adrenal tissue and high telomerase activity was associated with unfavorable prognosis (Hiyama et al. 1995; Poremba et al. 1999). The telomerase reverse transcriptase (TERT) gene encodes the protein constituent of telomerase. Rearrangements upstream of TERT were found to activate TERT expression in a subset of high-risk tumors (Peifer et al. 2015; Valentijn et al. 2015). These studies implicate telomere maintenance by ATRX loss and TERT activation as important mechanisms in metastatic and high-risk neuroblastoma. And they show that genomic aberrations may influence disease progression by modifying transcript structure or modulating the expression of their gene target.

Rare germline variants in the neurodevelopmental gene PHOX2B were found to predispose to neuroblastoma through investigation of familial neuroblastoma cases and a patient with both neuroblastoma and Hirschsprung disease, another malformation originating from neural crest cells (Trochet et al. 2004). Rare PHOX2B variants were later found in 6% of cases of suspected hereditary origin and these mutations interfered with terminal differentiation (Raabe et al. 2008). Genome-wide association studies (GWAS) identified common germline variation in non-coding regions to predispose to neuroblastoma; and this variation was associated with expression traits of genes in their vicinity (Pandey et al. 2014; Russell et al. 2015; Bosse et al. 2012; Diskin et al. 2009, 2012; D. A. Oldridge et al. 2015; McDaniel et al. 2017; Chang et al. 2017; John M. Maris et al. 2008; Capasso et al. 2009; K. Wang et al. 2011). These investigations implicate both rare coding and common non-coding germline variants in the development of neuroblastoma and suggest that non-coding risk variants can exert their effect by modulating expression of genes in their proximity.

2.2 Genetic and epigenetic regulation of gene expression

Gene expression is the mechanism by which an organism employs DNA to synthesize gene products. These products are RNA molecules and proteins, that are translated from messenger RNAs (mRNAs) of protein-coding genes. Gene expression is essential for an

organism to develop a phenotype from its genotype. Cells respond to developmental cues and environmental conditions by changes in gene expression and thus its regulation is tightly controlled. Mechanisms of gene regulation include gene dosage, the control of transcription (the synthesis of RNA from DNA), RNA processing, stability and degradation, and the control of translation (the synthesis of proteins from RNA). In a broad sense it also includes mechanisms controlling the function and life cycle of proteins, such as their processing, localization and degradation (Buccitelli and Selbach 2020). The regulation of transcription from DNA to RNA is essential to regulation of gene expression. Rates of RNA transcription and degradation constitute steady state RNA level and the abundance of mRNA is an important determinant of cellular protein levels (Schwanhäusser et al. 2011; Buccitelli and Selbach 2020). In fact, mRNA levels were used to predict protein levels and protein activities (Wilhelm et al. 2014; Edfors et al. 2016; Alvarez et al. 2016), despite controversies over the accuracy of estimating protein abundance from mRNA abundance alone exist (Fortelny et al. 2017; Wilhelm et al. 2017). In eukaryotes the DNA is transcribed to mRNA by the enzyme RNA-Polymerase II (Pol II). The amount of RNA produced by transcription depends on the availability of DNA template and the efficiency of Pol II binding and elongation. Both cis- and trans-acting factors regulate gene expression by controlling transcription.

Cis-regulation of gene expression is the control of RNA synthesis by local genetic and epigenetic factors at the transcribed locus. It is an important layer of transcriptional control in which the effect is confined to the transcribed allele (Wittkopp, Haerum, and Clark 2004; Gilad, Rifkin, and Pritchard 2008) and that is often defined by the effect of cis-regulatory elements. In contrast, trans-regulation is the control of expression of genes independent of the allele or genomic location of its target. Trans-regulation is conducted by regulatory signals, such as transcription factors that interact with cis-regulatory sequences (Wittkopp, Haerum, and Clark 2004). Transcription factors are diffusible and can thus bind to many different genomic regions. They are not constrained to regulate a target gene on a specific allele or in a specific part of a chromosome. Figure 2 shows a schematic representation of trans- and cis-regulation of gene expression.

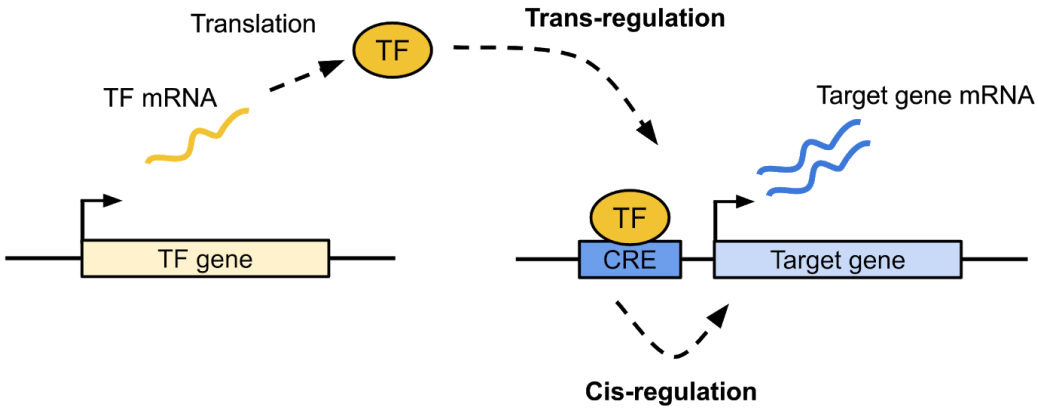


Figure 2: Scheme depicting cis- and trans-regulation of a target gene. Gene expression gives rise to a diffusible transcription factor (TF), that regulates a target gene in trans. Proximal to the target gene the factor binds to a cis-regulatory element (CRE), that regulates target gene expression in cis.

Cis- and trans-regulation of gene expression cannot be seen idependently: Cis-regulation can be interpreted as modulator of trans-regulation, resulting in different effects for the same trans-regulatory input signal. For example, a constant level of transcription factor activity can result in different rates of RNA transcription from the two homologous copies of a gene in the human genome. This may occur if sequence differences in regulatory regions of the gene result in unequal affinity in transcription factor binding. The overall transcriptional output is then a function of both cis- and trans-regulation across the two alleles. Sequence variants regulate genes both in cis and trans. Coding variants in the gene of a transcription factor can modulate the protein's regulatory signal and subsequently its targets in trans. Because the expression of genes encoding transcription factors are also regulated in cis, a cis-regulatory variant at such a locus may be responsible for a trans effect on the transcription factor's targets. Cis-regulatory variants that do not act in trans are generally expected to have limited effects on traits, because their effect is restricted to a small number of proximal genes. This is in contrast to trans-regulatory variants, that may frequently introduce pleiotropic effects (i.e. simultaneous effects on multiple phenotypes) due to a broader range of downstream consequences. For this reason cis- and trans-regulatory variants are expected to have qualitively distinct contributions in trait evolution (Wray 2007).

In the example of cis-regulation desribed above, two gene copies were available for transcription. Even though we generally find two copies per gene in the human genome (one on each of the two homologous chromosomes) there are exceptions: For example, the X chromosome in males is present only as a single copy. Additionally, many structural variants

were identified in the human genome, including deletions and duplications of chromosomal segments that increase or decrease the number of gene copies per individual (Iafrate et al. 2004; Sebat et al. 2004). In cancer, somatic alterations frequently change the copy-number of chromosomal regions that include the coding sequences of many genes (Beroukhim et al. 2010; Zack et al. 2013; Weischenfeldt et al. 2017). But how does gene dosage affect the expression level of RNAs? Several early studies addressing this question found a positive correlation between dosage and expression, but also found substantial differences in how strong individual genes are affected (Henrichsen et al. 2009; Schuster-Böckler, Conrad, and Bateman 2010; Jun Zhou et al. 2011). These findings demonstrate that in addition to cis-regulatory variation, copy-number variation is an important determinant of gene expression and that this form of genetic regulation is of particular importance in cancer.

To better understand how genes in neuroblastoma are regulated by genetic and epigenetic factors at their locus, I will focus on cis-regulation and copy-number-induced dosage effects on gene expression in this work. After having demarcated cis- from trans-regulation and briefly defined copy-number dosage effects, I will first introduce concepts of cis-regulation, including cis-regulatory elements, chromatin accessibility and epigenetic modifications in section 2.1.1. In section 2.1.2 I will then describe how genetic variation can alter cis-regulatory- and dosage effects on gene expression and give examples of cases where this form of genetic control has been associated with phenotypic consequences in the form of complex traits. Finally, in section 2.2.3 I will review somatic alterations and their role in the deregulation of gene expression in cancer.

2.2.1 Cis-regulatory elements and chromatin state

The vast majority of genetic information lies within non-coding regions, as protein coding genes only account for approximately 2% of the human genome. The non-coding genome harbors genetic elements that act as regulators of gene expression of nearby genes. These cis-regulatory elements (CRE) comprise promoters, enhancers and insulators. Promoters have an essential role in cis-regulation of gene expression, as Pol II binds to their DNA sequence. Each gene has a promoter that contains a transcription start site (TSS), which is the genomic position from which RNA synthesis starts in the 3’ direction of the coding DNA strand. Transcription by Pol II is controlled by transcription factors (TFs), proteins that can bind DNA sequence and facilitate or repress RNA synthesis. A set of TFs that bind to sequences at the core promoter constitute the pre-initiation-complex, a protein complex that regulates Pol II initiation and elongation (Kornberg 2007). Similar to promoters, enhancers

contain DNA sequences, to which TFs can bind. But in contrast to promoters, enhancers may be located several kilobases upstream or downstream of the TSS. TFs that bind to enhancers can recruit additional factors that may not bind DNA directly, but interact with Pol II and other factors (Thomas and Chiang 2006). Multiple enhancers and their bound factors and cofactors interact with a promoter to control transcription. Interactions between regulatory elements are mediated by chromatin contacts, loops of chromatin structure that allow distant regulatory elements to be close in the three dimensional space of the nucleus (Schoenfelder and Fraser 2019). Topologically associating domains (TADs) are contiguous regions in the genomes, in which contacts between regulatory elements occur more frequently (Pombo and Dillon 2015). Generally, transcription factors bound to enhancers are associated with transcriptional activation. Silencers are regulatory elements bound by repressive factors that attenuate gene expression, such as polycomb response elements, which are bound by transcriptional silencing factors of the polycomb family (Simon and Kingston 2009). Insulators are a class of regulatory elements that do not activate or repress transcription on their own, but instead limit the range of activating or repressing elements along the genome. The transcriptional repressor CCCTC-binding factor (CCTF) is a protein associated with insulators. Cohesion is a factor that mediates higher order chromatin structures by DNA looping and it co-localizes with CTCF at TAD boundaries (B.-K. Lee and Iyer 2012; Pombo and Dillon 2015). The interactions between enhancers and promoters allow for context-specific expression of genes. The more regulatory elements are involved in these interactions the more trans-regulatory signals can be integrated to control transcription in cis. These complex mechanisms are believed to have evolved to control gene expression in a highly tissue-specific manner throughout development (Heinz et al. 2015).

In the nucleus of eukaryotes DNA is wrapped around complexes of histones, proteins that act as spools for the DNA molecule. A histone complex consists of two subunits of histones H2A, H2B, H3 and H4 each. The nucleosome is a histone complex with its wrapped DNA. Together with other DNA-bound proteins nucleosomes consituate chromatin, the structure that makes up chromosomes. In heterochromatin nucleosomes are positioned close to each other, compacting the DNA and protecting it from interactions with other proteins. In contrast to heterochromatin, euchromatin contains loosely packed DNA, which is more accessible to other proteins apart from histones. In general, genomic regions that exhibit transcriptional activity are located in euchromatin. Many transcription factors are limited to bind accessible DNA sequences in histone-depleted regions of euchromatin. Other factors, so called pioneering factors, bind to closed chromatin in order to increase accessibility for factors that

subsequently bind to the region made accessible (Zaret and Carroll 2011). Thereby, CREs can be switched from an inactive to an active state, a process that underlies cell type differentiation during development (Velasco et al. 2017; Siersbæk et al. 2017; Rubin et al. 2017; Lopez-Pajares et al. 2017). Because accessibility of DNA sequence is seen as a prerequisite for active regulatory elements, accessible DNA regions ("open chromatin" regions) are investigated in order to predict promoters and enhancers in a cell-type and tissue-specific manner (ENCODE Project Consortium 2012; Shlyueva, Stampfel, and Stark 2014). Several assays were developed to measure chromatin accessibility genome-wide (Giresi et al. 2007; Schones et al. 2008; Song and Crawford 2010; Buenrostro et al. 2013). I will describe the open chromatin assay ATAC-seq in section 2.6.1.

Epigenetic modifications of histones characterize CREs and play an active role in gene regulation through chromatin remodeling and interaction with chromatin associated factors. The N-terminal tail of histones can be modified post-transcriptionally. The type of modification and its amino acid target are characteristic to the genomic region's functional activity. For example, H3 histones modified by acetylation of lysine at amino acid position 27 (abbreviated "H3K27ac") are found at active promoters and enhancers (Z. Wang et al. 2008; Creyghton et al. 2010). Additionally, active promoters are enriched in H3 histones modified by trimethylated lysine 4 (H3K4me3), whereas in active (and poised) enhancers the same lysine is predominantly monomethylated (H3K4me1) (Liang et al. 2004; Heintzman et al. 2007). Acetylation of histone tails weakens the interaction between histones and DNA and increases chromatin accessible to transcription factors (Allfrey, Faulkner, and Mirsky 1964; D. Y. Lee et al. 1993). H3K4me3 is recognized by the Pol II pre-initiation-complex and facilitates its recruitment to core promoters of active genes (Lauberth et al. 2013). HP1, a family of heterochromatin associated proteins involved in epigenetic gene silencing, was found to bind to the repressive histone mark H3K27me3 (Lachner et al. 2001). Chromatin remodeling factors that add and remove histone modifications play an important role in transcriptional regulation (Hyun et al. 2017; Marmorstein and Zhou 2014). For example, factors can regulate DNA accessibility of CRE by adding and removing acetylation of H3K27. Immunoprecipitation-based methods are used to study histone modifications in context of the DNA sequence (Robyr, Kurdistani, and Grunstein 2003; Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007). Here, antibodies capture histones carrying specific modification together with the bound DNA. The DNA sequence captured is analysed to identify genomic regions associated with specific histone modifications in a cell type of tissue sample. The NGS-based assay ChIP-seq will be described in section 2.6.1.

Taken together, chromatin accessibility and histone modifications are investigated to describe chromatin state and study its association with transcription. Chromatin accessibility of cis-regulatory, such as promoters and enhancers, is required to control gene activity. Histone modifications control transcriptional through chromatin structure and interactions with chromatin-associated factors. Thereby histones take an active part in transcriptional regulation in cis through their epigenetic modifications. Assays that determine chromatin accessibility and histone modifications in the context of the DNA sequence are used to predict regulatory elements. Figure 3 shows examples of chromatin accessibility and histone marks associated with different states of CREs (Shlyueva, Stampfel, and Stark 2014).

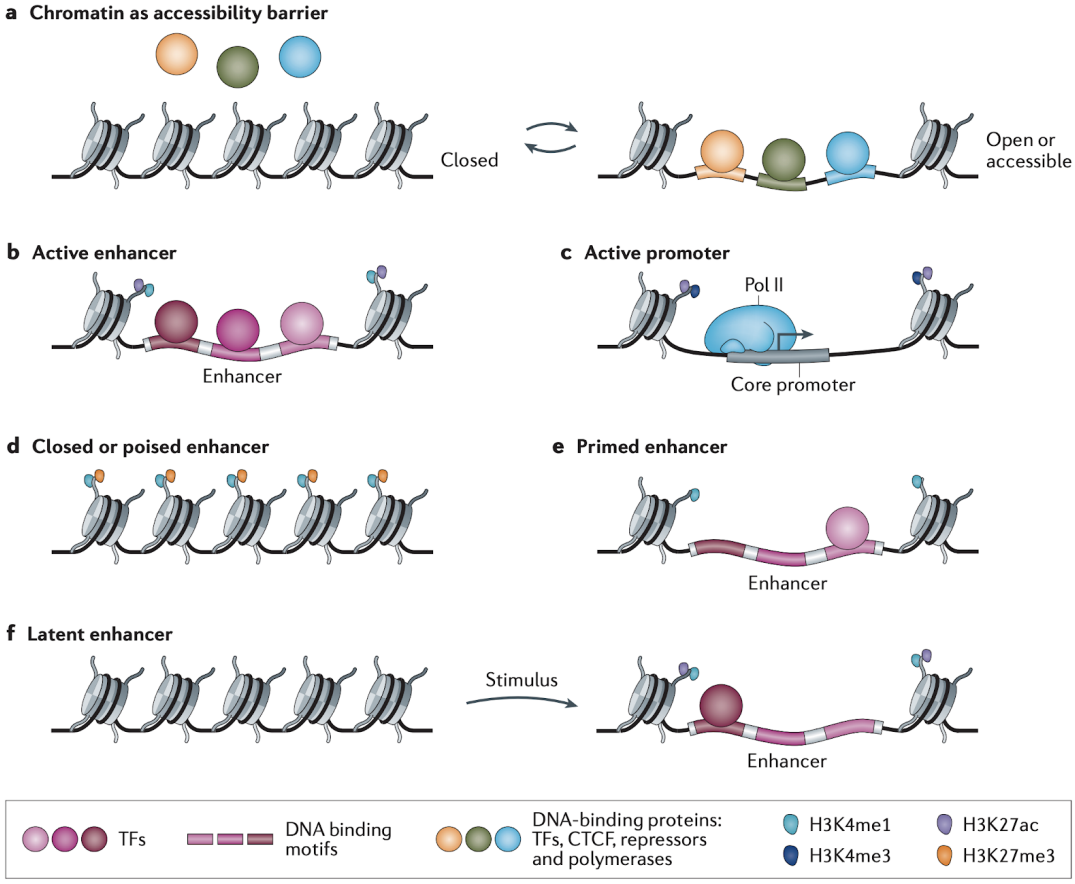


Figure 3: Chromatin accessibility and histone modifications at cis-regulatory elements. **a,** DNA-binding proteins bind to motifs in open but not closed chromatin. **b-c**, Transcription factors and Polymerase II (Pol II) bind active enhancers and active promoters respectively. Active promoters and enhancers carry characteristic histones modifications. **d**, Closed chromatin and repressive H3K27me3 mark at closed or poised enhancers. **e**, Primed enhancers marked by H3K4me1. **f**, Latent enhancer in closed heterochromatin lacks histone modifications, but becomes accessible on external stimulus and histones acquire active enhancer marks. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics (Daria Shlyueva, Gerald Stampfel, Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272–286, 2014), copyright Macmillan Publishers Limited (2014).

Besides methylation of histone tails, the methylation of cytosines in DNA constitutes an additional layer of epigenetic control in transcription. Here, methyl groups are added to cytosines of CpG dinucleotides. Methylated cytosines have an increased chance to be converted to thymine by deamination, and for this reason CpG dinucleotides are depleted in the human genome (Coulondre et al. 1978; Bird 1980). Those CpGs that remain are evolutionary conserved and form clusters known as CpG-islands. CpG are often found at the promoter and methylation of CpG islands inside gene bodies is associated with transcriptional silencing (Saxonov, Berg, and Brutlag 2006; Deaton and Bird 2011). Non-methylated promoters are in a transcriptionally permissive state. Methylated promoters may repress transcription by direct inhibition of transcription factor binding to methylated sequences or by recruitment of factors containing methyl-binding domains, which in turn recruit repressive cofactors (Deaton and Bird 2011). Despite the established relation between promoter methylation and transcriptional repression, generally CpG methylation can display both positive and negative correlation with transcription. The directional effect is best explained by histone marks and chromatin accessibility of differentially methylated regions (Wagner et al. 2014), indicating that the effect of methylation is specific to the chromatin context. The association with histone marks suggests that the characteristics of CREs may underlie the directional effect. E.g. methylation of an enhancer could attenuate expression of a target gene, while methylation of a repressor could increase its expression.

Methylation of DNA and histones underlie genomic imprinting. Genomic imprinting (or simply “imprinting”) is an epigenetic control mechanism found in plants and animals. It involves the transcriptional silencing of the paternal or maternal copy of a gene. As a consequence, only one of the two alleles is expressed. In mammals, imprinting plays an important role in gene regulation during prenatal development and deregulation of imprinted loci were found to cause a variety of developmental syndromes (Peters 2014). In the human genome, imprinted genes are often located in clusters. Many orthologues imprinted gene clusters exist between human and mouse and therefore mouse models have been used to study elements that control imprinting of human disease-associated gene clusters, such as the GNAS, IGF-H19 and DLK1-DIO1 locus (Williamson et al. 2004; Fröhlich et al. 2010; Sun et al. 1997; da Rocha et al. 2008). Gene clusters contain cis-acting imprinting control regions (ICR)s, that establish parent-specific expression. ICRs exhibit patterns of DNA methylation and histone modifications specific to the parental origin of the allele (Williamson et al. 2004; Henckel et al. 2009). Unmethylated ICRs are active and associated with a permissive chromatin state. Active ICRs can silence protein coding genes by inducing expression of a

non-coding antisense transcript. For example, in the Gnas gene cluster in mice an ICR at the promoter of a non-coding transcript antisense of Nesp silences Nesp expression from the paternal allele in cis (Williamson et al. 2011). Active ICRs may also silence target genes through the recruitment of repressive factors. For instance, the maternal allele of the Igf-H19 gene cluster in mice contains an ICR that recruits the repressive factor CTCF and transcription of Igf2 is then silenced through CTCF-induced insulation of Igf2 from downstream enhancer elements (Bell and Felsenfeld 2000; Hark et al. 2000). Imprinting is neither static nor absolute, as it may develop in the course of embryonic development and gene expression from imprinted alleles may still be detected after imprinting occurred (Latos et al. 2009; Sasaki et al. 1992). Imprinting or repressive methylation in general conveys a dosage effect on gene expression in which higher methylation levels and stronger imprinting are associated with decreased gene expression levels from the affected allele. Together, these findings show that imprinting is an important process in the control of the developmental gene expression program. It is controlled by ICRs, regulatory elements that when active facilitate the imprinted state of gene clusters through different downstream control mechanisms in cis, such as antisense transcription and enhancer insulation.

Im summary, cis-regulation is orchestrated by the interaction of non-coding CRE and epigenetic modifications. TFs bind to these elements to enhance or repress transcription. TADs control interactions of regulatory elements by establishing isolated neighborhoods of genes and CREs. Nucleosomes are the envelope of DNA and how densely they are packed determines the degree of chromatin accessibility. Active CREs reside in accessible (“open”) chromatin regions and can be bound by TFs. Histones are the molecular building blocks of nucleosomes that are modified post-transcriptionally. These epigenetic modifications are characteristic to the functional activity of the associated genomic sequence and several combinations of modifications have been associated with classes and states of CREs. DNA methylation of cytosines in CpG dinucleotides constitutes an additional layer of epigenetic control, which is often found at promoters but also at other CREs. DNA methylation underlies genomic imprinting of gene clusters, an important regulatory mechanism in prenatal development. Imprinting is controlled by specific CREs (namely ICRs) and is associated with mono-allelic expression of imprinted regions in a parent-of-origin-dependent manner.

2.2.2 Genetic variation in gene expression

Genetic variation is responsible for differences in cis-regulation between alleles and individuals. Variation can modulate the effect of regulatory elements by sequence changes

to DNA motifs affecting binding affinity of DNA binding proteins, such as transcription factors and readers and writers of DNA methylation. Studies aiming to identify cis-regulatory variation reported an enrichment of expression-trait associated variants at the TSS of genes (Stranger et al. 2005; Veyrieras et al. 2008; Stranger et al. 2012), indicating that variants proximal to the promoter have strong potential cis-regulatory effects. A study on SNPs and gene expression in yeast revealed that genetic variation in motifs of transcription factor binding sites (TFBS) at promoters are predictive for expression differences between yeast strains (K. Chen et al. 2010). Studies of trait-associated variation located in enhancers pinpoint SNPs that are responsible for differences in long-range cis-regulation between individuals. Enhancer variation linked to obesity and type 2 diabetes was found to regulate expression of a body-mass implicated gene at ~400kb TSS distance (Smemo et al. 2014). Enhancer variation common in the European population modulates expression of a target gene associated with blond hair color at ~350kb TSS distance in hair follicle cells (Guenther et al. 2014). Guenther and colleagues showed that the regulatory variant caused strong phenotypic differences in the hair color of mice despite reducing enhancer activity by only 22% in vitro (Guenther et al. 2014; Corradin and Scacheri 2014), indicating that moderate differences in regulatory control can have strong phenotypic consequences. Similarly, common variation linked to fetal hemoglobin levels in intronic open chromatin was found to be associated with modest changes in binding differences of transcription factors GATA1 and TAL1, and is associated with expression of its target gene specifically in erythroid cells (Bauer et al. 2013). Massively parallel testing of promoter and enhancer sequences by functional reporter assays has helped to quantify the regulatory potential of individual sequences (Melnikov et al. 2012; Patwardhan et al. 2012; Arnold et al. 2013; Sharon et al. 2012; Mogno, Kwasnieski, and Cohen 2013). Assays based on exhaustive mutagenesis of individual sequences quantified regulatory differences at single nucleotide resolution, revealing the potential functional impact of base pair substitutions: A reporter assay experiment in mouse liver covering all possible substitutions in three mammalian enhancer sequences determined that most variants have modest effect on enhancer activity, indicating that enhancers may be robust against most sequence changes (Patwardhan et al. 2012). However, variants that show strong differential enhancer activity in reporter assays co-localize with known motifs, demonstrating that enhancer elements are specifically sensitive to sequence changes at TFBSs (Patwardhan et al. 2012; Melnikov et al. 2012). Synthetic sequences were used in reporter assays to examine the functional consequence of CRE composition in promoters (Sharon et al. 2012; Mogno, Kwasnieski, and Cohen 2013) and these studies identified transcriptional activity was influenced by CRE multiplicity,

orientation and distance to the core promoter as well as cooperative binding of transcription factors.

Variation in CTCF binding affect cis-regulation of genes by switching long-range promoter-enhancer interactions. Cis-regulation by distal elements depends on higher order chromatin organization, which is in part mediated by CTCF binding in insulator regions (Ong and Corces 2014; Pombo and Dillon 2015). Thus, its is expected that variation in CTCF binding sites can impair insulator functionality and give rise to patterns of chromosomal contacts with specific cis-regulatory interactions. CTCF stabilizes interactions between promoters and enhancers and disruption of CTCF binding motifs or CTCF depletion was found to cause increased variability in gene expression (G. Ren et al. 2017). Variants in CTCF binding sites have been associated with increased risk for breast cancer (Dai et al. 2015) and differential binding of CTCF was attributed to genetic differences clustered around CTCF binding sites (Ding, Ni, et al. 2014). Two asthma risk loci were found to disrupt CTCF binding and consequently upregulate ORMDL3, a negative regulator of interleukin, by establishing contacts between distal CREs and its promoter (Schmiedel et al. 2016). In gastrointestinal tumors disruption of CTCF binding sites by mutational hotspots was found to be associated with changes in expression of genes within the TADs adjacent to the affected CTCF binding sites (Guo et al. 2018). These studies show that variation in CTCF elements can have cis-regulatory effects associated with disease phenotypes. And they suggest that variation within these elements may be able to reorganize enhancer-promoter interactions due to differential binding affinity of CTCF with downstream consequences on insulation and chromatin looping.

Variation in regulatory regions is expected to exert cis-regulatory effects by changes in transcription factor binding. Additionally, regulatory regions are often found in accessible chromatin. Consequently, variants inside open chromatin regions are promising candidates for functional regulatory variation. However, because open chromatin regions can be large and not all variants inside these regions co-localize with TFBS, the mere observation that a variant is accessible is not sufficient evidence for a functional role. Additional evidence can be collected by associating variant genotypes with quantitative traits and these analysis strategies will be introduced in section 2.6.4. A comparison of variant-associated changes in chromatin accessibility and gene expression revealed that the loci associated with these traits are linked (Degner et al. 2012). Specifically, Degner and colleagues found variants that affect DNAse-I-hypersensitive sites (DHS) to be located within or very close to these sites

and frequently modulate transcription factor binding. Additionally, DHS-associated variants were found to be enriched for associations with gene expression differences in cis, an effect that was stronger the closer the variant was to the gene's TSS (Degner et al. Figure 4a). Besides chromatin accessibility, genetic variation can additionally be associated with other chromatin traits, such as histone marks and transcription factor binding.
Chromatin-associated variants were found to have coordinated effects on multiple local molecular traits, including H3K27ac, H3Kme1, H3Kme3 and gene expression (Grubert et al. 2015; Waszak et al. 2015). Chromatin contact mapping revealed that chromatin traits are coordinated in local structures inside TADs and three dimensional contacts between regulatory elements underlie distal chromatin trait associations (Grubert et al. 2015). Local chromatin traits were found to be highly correlated and chromatin trait-associated variation enriched for transcription factor occupancy (Waszak et al. 2015). Similar to Degener et al., Waszak and colleagues found that despite similar effect on chromatin traits, associated variants proximal to TSSs have stronger effect on gene expression and they conclude that promoter variants have either inherently larger effects on gene expression or the effect of enhancer variants is highly context-specific (e.g. dependent on a stimulus). Together, these studies suggest that functional variants with effect on chromatin accessibility and histone marks co-localize with regions of their respective traits. Variants in promoters may be the strongest cis-regulators, as variants close to TSS are more likely to cause gene expression differences. Potentially functional variants correlate with regionally clustered chromatin traits inside TADs and intergenic chromatin-associated variants likely affect enhancer activity but have generally weaker direct effects on gene expression. A possible explanation is a smaller effect size of single enhancers in the control of gene expression of target genes. Additionally, these reports indicate that prioritization of variants close to open-chromatin regions and transcription start sites may guide the identification of functional cis-regulatory variation.

Cis-regulation by DNA methylation is affected by genetic variation at or proximal to regulatory elements. Studies investigating variant consequences for DNA methylation and gene expression are beginning to reveal regulatory interactions between these molecular phenotypes, genotypes and complex traits. Genetic variation linked to differential DNA methylation is recurrenlty associated with gene expression differences in cis. A study in 62 individuals reported that 9.5% of regions with variation-associated gene expression differences contain variation that was associated with both expression and DNA methylation (Wagner et al. 2014). Variants can influence gene expression through changes in DNA methylation, a mechanism that underlies complex trait associations. In an examination of

variant-mediated effects of DNA methylation and expression on complex traits, both quantitative traits showed substantial overlap (31.4%) in genes associated with 17 complex-traits investigated (Hannon et al. 2017). Furthermore, in this study Hannon and colleagues found that DNA methylation-linked variation is more likely to be associated with local gene expression differences than variation that does not affect methylation, underlining the importance of DNA methylation in transcriptional control by regulatory variation. Conserved TFBS are enriched in CpG island and hypomethylated CpG islands discriminate true binding sites from those computationally predicted but unbound by their respective factor (Choy et al. 2010), indicating that many elements controlling transcription may require a sequence environment that is responsive to DNA methylation. Variants associated with DNA methylation are enriched in open chromatin and repressive histone marks and depleted in active chromatin marks (Hannon et al. 2018), showing that variation affects methylation levels at regulatory elements, an interaction that is particular prevalent in the context of transcriptional repression. Functional genetic studies in mice identified ICRs in differentially methylated regions of imprinted gene clusters and showed that mutations of these regions result in loss of imprinting. Deletion of an ICR at the Gnas locus in mice abrogates the tissue-specific paternal imprinting of Gnas and leads to its bi-allic expression (Williamson et al. 2004). Mutation in CTCF target sites of ICRs at the H19 locus results in loss of imprinting at the IGF2 locus (Pant et al. 2004) and abrogating mutations in CTCF and OCT binding sites at this locus are associated with Beckwith-Wiedemann syndrome (Sparago et al. 2004; Prawitt et al. 2005; Demars et al. 2010). Taken together, these reports imply that DNA methylation and gene expression is under control by genetic variants and integrative analyses established functional links between these molecular phenotypes. Furthermore, they suggest that regulatory elements, that control gene expression in cis, are themselves under control by DNA methylation. It is expected that some functional variants in regulatory elements may not affect transcription directly (e.g. through differential binding of factors interacting with the transcriptional machinery), but rather affect this process indirectly by causing differences in methylation at CREs. Despite possibly being less direct, these sequence changes can have profound downstream consequences, evident by disease-associated variation in ICRs.

Copy-number variation introduces DNA dosage-dependent regulation of gene expression. Besides SNPs, frequently occurring genetic variation also include segmental copy-number variants (CNVs), and their discovery challenged the idea of a common human genomic reference, with only minor base-pair differences between individuals (Iafrate et al. 2004;

Sebat et al. 2004). CNVs can modulate gene expression by changes of gene dosage and this effect is traditionally seen independent from cis-regulation, as dosage changes often span entire sets genes including their regulatory elements. Our definition of cis-regulation requires the regulatory effect to be constrained to a single allele, but CNVs are not necessarily confined to one allele, as e.g. a duplicated gene copy can be located on a different chromsome than the original copy. However, CNV dosage effects do not satisfy the definition of trans regulation either, as they are not mediated by a diffusible cis-acting factor. Furthermore, the dosage-dependent CNV effect is limited to the DNA sequence it encompasses, and is therefore regionally constrained, a property that makes it similar to cis-regulatory effects. As CNV dosage effects are related to cis regulation by their regional constrained and as somatic CNVs (also referred to as SCNAs) are abundant in cancer genomes I will include this class of genetic control of regulation here and later also examine its local effects on gene expression in neuroblastoma.

CNVs may be the cause of differences in gene expression between individuals, and smaller variations, such as SNPs in LD with the causal CNVs may be mistaken for cis-regulatory SNPs (Dermitzakis and Stranger 2006). CNVs were found to explain 17.7% of the genetic variation in gene expression and the contribution was largely mutually exclusive to those explained by SNPs (Stranger et al. 2007). In an early study on CNVs and expression in different tissues across mice strains CNVs-affected genes showed a higher variance in gene expression across tissues and individuals, were enriched for differentially expressed genes and their expression was significantly positively correlated with DNA abundance (Henrichsen et al. 2009). Similarly, studies in drosophila and human lymphoblast cells found that expression of genes in CNVs was up- and down-regulated in increases and decreases of CN respectively (Schuster-Böckler, Conrad, and Bateman 2010; Jun Zhou et al. 2011). However, the copy-number induced effect on gene expression was smaller than expected by proportional effects of DNA dosage (Schuster-Böckler, Conrad, and Bateman 2010) and substantial heterogeneity in dosage sensitivity was observed across genes (Schuster-Böckler, Conrad, and Bateman 2010; Jun Zhou et al. 2011). Different mechanisms underlying expression changes that may be observed for gene duplications were suggested (Henrichsen, Chaignat, and Reymond 2009): In the simplest scenario additional gene copies induce a proportional increase of RNA due to a copy number dosage effect. Interactions between the duplicated gene and its chromatin context, CREs or trans effects could explain changes in gene expression that are dis-proportional to the copy-number increase: Negative feedback mechanisms of the gene product on its own transcription could lower RNA levels.

Similarly, dosage-proportional increases of a target gene could be attenuated by negative regulation of an early expressed repressors residing on the same duplicated segment. Regulatory elements that are important in transcriptional regulation of the target gene could be missing on the gained segment, resulting in a lower expression from the gained segment specifically. Non-tandem duplications may reside on different chromosomes and the additional copy may therefore be embedded in a “foreign” chromatin context that prevents transcription of the copy, despite availability of endogenous cis-regulatory elements in cis. Finally, tandem duplications may impact chromatin structure and prevent efficient transcription from either of the copies.

Gene dosage is expected to be the underlying cause of pathogenesis in many CNV disorders such as 22q11.2 deletion syndrome, Williams-Beuren syndrome, Prader-Willi syndrome, Charcot-Marie-Tooth disease type 1A/hereditary neuropathy with liability to pressure palsies and sotos syndrome (Shaikh 2017). Disease CNVs are often large and it is difficult to pinpoint individual genes whose CNV-induced expression change is responsible for the phenotype. However, some studies have indeed linked CNV-induced dosage effects on gene expression of individual genes with disease phenotypes: CNVs of the HBD-2 gene were found to be associated with Crohn’s disease and low copy numbers of the gene showed decreased mRNA levels in colonoscopy biopsies (Fellermann et al. 2006). A duplication of the TLR7 gene in mice is associated with systemic lupus erythematosus and the duplicated locus is found to induce higher expression levels of TLR7 in B cells (Pisitkun et al. 2006). Segmental duplications affecting the CCL3L1 gene were found to be associated with HIV/AIDS susceptibility and CCL3L1 levels were associated with its gene CN (Gonzalez et al. 2005). These reports show that CNVs are common in the human genome and can alter expression of affected genes in a dosage-dependent manner. Individual CNVs may thereby underlie the pathogenesis of CNV-linked disorders. Dosage-dependent regulation of gene expression by CNVs is different from cis- and trans-regulation but resembles cis-regulation in that it is constrained to its genomic region. Many genes do not show proportional changes of expression when affected by CNVs and are thus insensitive (or less sensitive) to DNA dosage. These differences in sensitivity indicate compensatory mechanisms counterbalancing effects of CNVs.

Genomic rearrangements can alter the regulatory context in a dosage-independent manner, causing deregulation of genes in their vicinity. CNVs were found to induce expression differences of genes outside of the region affected by the copy-number change. For

example, genes proximal to a deletion on chromosome arm 7q in Williams-Beuren syndrome displayed patterns of deregulation, with stronger effects closer to the deletion breakpoint (Merla et al. 2006). Similarly, an engineered duplication on chromosome arm 17p in a Potocki-Lupski syndrome mouse model displayed altered expression patterns for genes not only within but also flaking the engineered interval (Molina et al. 2008). A comparative analysis of copy-number and transcriptome between six tissues in three mice strains found genes between 50-250 kb from CNV breakpoints to have significantly higher expression variance compared to distal genes with effects on genes up to 450 kb from the breakpoint (Henrichsen et al. 2009). These findings identified dosage-independent regulatory effects of large variation on gene expression and pointed to CNV-induced structural changes as an underlying cause for gene deregulation. Structural variation (SV) may occur with CN changes, such as deletions, duplications and gains beyond an additional copy, including amplifications (as found in many cancers). But SVs may also be CN-neutral in the case of inversions or more complex rearrangements that connect distal parts of chromosomes or even different chromosomes. Next generation sequencing technology (Section 2.6.1) allowed the development of methods that identify SV independent from CN changes (Mahmoud et al. 2019). Many recent methods can reveal targets and corresponding breakpoints of rearrangements at base pair resolution. Sequence-based methods applied to human samples from several tissues implicated a substantial contribution of common SVs to gene expression differences and found strong enrichment of rare SVs proximal to gene expression outliers (Chiang et al. 2017). The SV-induced effect on gene expression was found to be positively associated with SV length and functional SVs were enriched in those with chromatin contacts to the promoter of the target gene, where variants closer to the promoter-distal looping anchor contributed more lead associations (Jakubosky et al. 2020). The mechanisms by which SVs affect gene expression largely depend on the sequence- and chromatin context (Spielmann, Lupiáñez, and Mundlos 2018): SVs inside TADs that do not interfere with insulating elements can alter the dosage of cis-regulatory elements in contact with the promoter of a target gene within the same chromatin domain, resulting in higher or lower levels of expression dependent on quantity and distance between CRE copies and the gene. In contrast, SVs between TADs may cause changes in higher-order chromatin structure, creating ectopic loops and contacts between regulatory elements that would otherwise not be formed. Such SVs may fuse neighboring TADs to bigger domains by removal of insulator elements or create de-novo TADs, insulating smaller regulatory environments of genes, that would otherwise be in contact with different or additional CREs.

Non-coding SVs are associated with mendelian and rare disease and these variants were found to induce changes in gene expression by re-programming interactions between promoters and distal regulatory sequences. Non-coding intronic SVs in the form of repeat expansions are genetic determinants of the neurodegenerative Parkinson's (Schüle et al. 2017) and Huntington's disease (McColgan and Tabrizi 2018) and complex SVs were identified in cases of rare mendelian disorders (Sanchis-Juan et al. 2018). SV were implicated in transcriptional deregulation of the TAF1 gene in X-linked dystonia-parkinsonism (Bragg et al. 2017). Recent studies found SVs to be implicated in the shuffling of chromatin domains of developmental genes causing disease phenotypes. Copy-number neutral inversions at the Wnt6-Pax3 locus in mice were found to disrupt CTCF boundary domains and alter gene expression by rewiring promoter-enhancer interaction, leading to abnormal limb development in mammals (Lupiáñez et al. 2015). Duplications of an enhancer array proximal to the Ihh gene in mice are associated with variable expression of Ihh proportional to the number of enhancer copies and their distance to Ihh resulting in several developmental defects (Will et al. 2017). In Cooks syndrome an aberrant chromatin context is formed by a TAD boundary element duplication, which enforces ectopic interactions between Sox9 regulatory elements and the Kcnj2 gene causing its misexpression (Franke et al. 2016). And a deletion discovered in patients with autosomal dominant adult-onset demyelinating leukodystrophy causes ectopic adoption of at least three enhancers to the LMNB1 promoter resulting in its overexpression (Giorgio et al. 2015). Together, these findings show that SVs can regulate gene expression in cis independent of gene dosage through DNA sequence changes that affect the order of regulatory elements. SVs include CNVs, such as deletions and duplications, because losses or gains cause rearrangement of DNA sequences. CN-neutral SVs, such as inversions, only affect the order of DNA sequences but not dosage of genes or regulatory elements. Studies linked non-coding SVs to increased gene expression variance and positional effects indicate that SVs can alter the regulatory context of proximal genes. Studies of developmental and disease phenotypes revealed interactions between SVs and TADs in the control of gene expression in cis. SVs constrained to single TADs may alter dosage and distance of regulatory elements relative to a target gene's promoter within its entopic regulatory environment. SVs affecting insulator elements shuffle interactions between promoters and enhancers of neighboring TADs, and may create ectopic interactions between regulatory elements that would otherwise not form. SV induced cis-regulatory effects thereby play a crucial role in cis-regulation of gene expression.

Genetic elements and the epigenetic state constitute a framework for the control of gene expression through transcriptional regulation. The function of regulatory elements, such as promoters, enhancers, insulators and ICRs depend on their genomic sequence and their interaction. Consequently, variation affecting sequence, position and dosage of regulatory elements can alter their function. For example, a sequence variant can weaken or even abolish the interaction between a regulatory element and a DNA-binding factor. Or variants may introduce new sequence motifs, that are preferentially bound by different DNA-binding factors with distinct regulatory consequences. Single nucleotide variants (SNVs) and smaller multi-nucleotide variants (MNV) may change regulatory motifs. Additionally, copy-number variation may alter gene expression by changes to the number of regulatory elements (e.g. duplication of an enhancer), or simply in a dosage-dependent manner by increasing the DNA template available for transcription. Here, gained or lost copies of genes may include both the coding sequence and their regulatory elements. SVs in non-coding regions can alter gene regulation by rearrangements of the DNA sequence that change interactions of regulatory elements and their effect on target genes. Taken together, multiple forms of genetic variation impact gene regulation by changes in cis-regulatory control or by dosage effects.

2.2.3 Deregulation by somatic alterations in cancer

Cancer is a disease of the genome, in which aberrant genetic changes drive processes that lead to specific traits, including increased proliferation (through self-sufficient growth and insensitivity to anti-growth signals) and replicative immortality (D. Hanahan and Weinberg 2000). Cancer genomes harbor a variety of somatic alterations (also termed mutations), genetic variants that do not stem from the germline but that have accumulated in the somatic tissue. Somatic alterations found in cancer include SNVs, small insertions, deletions and substitutions, SVs and somatic SCNAs. Somatic SNVs are basepair-size insertions, deletions or substitutions, and the smallest class of somatic alterations. Small insertions, deletions and substitutions comprise a group of medium-sized alterations from two to several base pairs in size. Similar to germline SVs, somatic SVs are characterized by their copy-number effect, such as duplication or deletion and the structural changes they introduce relative to the reference genome sequence. These also include copy-number-neutral events, such as inversions and translocations. SVs often fall together with copy-number alterations and in cancer genomes patterns of co-occurrence characterize high-level SV classes (Y. Li et al. 2020). In contrast to germline SVs, structural variants in cancer frequently connect even distal genomic regions and regions from different

chromosomes (inter-chromosomal rearrangements) (Baca et al. 2013; Hasty and Montagna 2014; Cortés-Ciriano et al. 2020). SCNAs include deviations in chromosomal copy number, chromosome arm-sized or smaller segmental and focal (< 3Mb) copy-number changes. SCNAs are large variants that introduce DNA copy-numbers different from the germline genome, where generally two copies per homologous region are present for the 22 autosomes in both sexes. In that respect SCNAs are analogous to germline CNVs. However, in contrast to CNVs, SCNAs in cancer occur at high frequency and recurrently affect whole chromosomes and chromosome arms resulting in aneuploidy (Zack et al. 2013; Ben-David and Amon 2019). It is expected that gained DNA copy-number segments must translocate to chromosomes or form extrachromosomal circular DNA (Section 2.5) in order to stably replicate and segregate into daughter cells during mitosis. Additionally, segmental losses require ligation of non-contiguous DNA segments along the affected chromosome. Thus, segmental SCNAs coincide with somatic SVs, inherently linking these two variant classes.

By cell divisions somatic alterations are transmitted to daughter cells. Somatic alterations that drive proliferative mechanisms are termed driver mutations. Mutations without such consequences are known as passenger mutations. Initially, mutations are limited to the single cells in which they occur, but those cells that obtained growth advantages by driver mutations show stronger proliferation and can expand in the population of malignant cells, a process termed clonal evolution (Nowell 1976). Here, “clonal” indicates that the evolving population consists of clones of a common progenitor cell and is not the product of recombination between individuals (as is the case in the evolution of species). Cancer genes are genes with convergent patterns of mutations within single- or across different cancer types (Bamford et al. 2004). Cancer genes that confer growth advantage when activated are termed oncogenes (or proto-oncogene to emphasize their physiological role in healthy tissue or when not activated by driver mutations). Similarly, cancer genes that confer a growth advantage when deactivated are termed tumor suppressor genes. Early studies on somatic alterations in cancer focused on somatic SNVs in exons of protein-coding genes in order to identify mutations in oncogenes and tumor suppressor genes. Such mutations can alter the structure of proteins translated from the resulting RNA, thereby changing the protein function. For example, a somatic mutation resulting in an amino acid change of the translated protein can constitutively activate a functional protein domain of a proto-oncogene. Similarly, deleterious mutations in coding sequences impair protein function by disrupting functional domains, a phenomenon associated with tumor suppressor genes. With recent advances in WGS somatic driver alterations were also found in the non-coding

tumor genome. Such driver mutations do not change the coding sequence of cancer genes but instead modulate their regulatory environment, resulting in deregulation of oncogenes and tumor suppressor genes. Genetic control of gene expression in tumor cells is the result of both the germline regulatory context and its modification by somatic regulatory drivers. In the following paragraphs I will highlight mechanisms by which somatic alterations of different variant classes control gene expression and give examples of such regulatory drivers previously identified across different cancer types.

Small and medium-sized somatic alterations, such as SNVs, insertions, deletions and focal amplifications of regulatory elements can cause deregulation of genes targeted by affected elements. Highly recurrent somatic SNVs in two distinct positions of the TERT promoter were first discovered in melanoma and these mutations were found to create de-novo binding motifs for ETS transcription factors associated with increased transcriptional activity (Horn et al. 2013; Franklin W. Huang et al. 2013). TERT promoter mutations were also found to be prevalent in glioblastomas, liposarcomas, oligodendrogliomas, bladder cancer and other types of cancer originating from tissues with lower rates of self renewal (Killela et al. 2013), indicating that activation of TERT telomerase maintenance by regulatory drivers is an important mechanism to escape cell senescence in these cancers. Non-coding TERT promoter mutations are exceptional in their recurrence and effect size on gene expression across 14 cancers examined (Fredriksson et al. 2014). In their study Fredriksson and colleagues also found candidate regulatory mutations in PLEKHS1, DPH3 at lower frequencies but could not associate them with expression levels of these genes. In contrast, somatic mutations in the 5’ UTRs of NFKBIZ found in 14% of activated B-cell diffuse large B-cell lymphoma showed effects on RNA expression levels (Arthur et al. 2018). A study of non-coding drivers in a large pan-cancer cohort confirmed regulatory mutations in the TERT promoter and NFKBIZ UTR and identified new candidate regulatory mutations in the promoter of MTG2 as well as UTRs of TP53 and TOB1 (Rheinbay et al. 2020). However, in their study Rheinbay and colleagues also pointed out that the number of regulatory point mutations discovered across 38 cancer types was unexpectedly low. Small insertions in enhancer elements are prevalent in cancer cell lines and patient samples (Abraham et al. 2017) and in T-cell acute lymphoblastic leukemia alterations of two distinct enhancers introduce binding motifs for transcription factor MYB causing overexpression of the TAL1 and LMO2 oncogenes respectively (Mansour et al. 2014; Abraham et al. 2017). In acute myeloid leukemias a cluster of distal enhancers that establishes contact to the MYC gene in leukemia cells was found to be focal amplified (Shi et al. 2013). In chronic lymphocytic

leukaemia somatic mutations in an enhancer element causes reduced expression of PAX5 (Puente et al. 2015). Copy-number analysis of non-coding enhancers identified in cell lines revealed that somatic gains of enhancer elements are associated with overexpression of KLF5, USP12, PARD6B and MYC in epithelial cancers (X. Zhang et al. 2016). Similarly, BRD4 and NOTCH3 expression is attenuated in breast and ovarian cancers harboring an intronic microdeletion of a potential enhancer element (Rheinbay et al. 2020).

Growing evidence suggests that SVs implicating regulatory elements affect regulation of genes in tumors. An early study in 1975 found a characteristic 8q-15q translocation in burkitt lymphoma cells (Zech et al. 1976) and later it was shown that these translocations connect the immunoglobulin (IG) heavy chain locus with the MYC gene (Dalla-Favera et al. 1982; Taub et al. 1982), implying the de-regulation of MYC by regulatory elements from the IG heavy chain locus. Similarly, in 30-40% of diffuse large-cell lymphoma cases genomic rearrangements juxtapose the BCL6 coding sequence to distal promoters, leading to overexpression of BCL2 (Ye et al. 1995). In CD3 T-cell acute lymphoblastic leukemias gene expression of TAL1 gene is put under control of the STIL (SIL) promoter by deletion of the STIL gene and its downstream intergenic region (Aplan et al. 1990; Breit et al. 1993). Similarly, in prostate cancer the gene TMPRSS2 is recurrently translocated to the 5’ ends of ERG or ETV1 resulting in overexpression of these ETS family transcription factors in response to androgen, an effect suggested to arise from androgen response elements in the TMPRSS2 promoter (Tomlins et al. 2005). In acute myeloid leukemia a GATA2 distal hematopoietic enhancer is translocated to the EVI1 gene, causing its overexpression and induction of EVI1 expression by this enhancer element induced neoplasms in a transgenic mouse model (Yamazaki et al. 2014). In medulloblastoma proto-oncogenes GFI1 and GFI1B are recurrenly activated by translocations of regulatory elements from either local chromosomes (regularly involving enhancers of the PRRC2B-DDX31 locus) or from distal chromosomes (Northcott et al. 2014). Northcott and colleagues found that the translocated regions harbor strong epigenetic marks characteristic for potent enhancer elements. In adenoid cystic carcinoma enhancer translocations cause overexpression of the MYB gene and chromatin conformation capture confirmed that the translocated elements interact with the MYB promoter (Drier et al. 2016). Interestingly, Drier et al. found that MYB itself binds to the translocated elements and thus the rearrangements may establish a short positive feedback loop of MYB auto-regulation.

Alteration affecting insulating CTCF elements at TAD boundaries could be a common scheme underlying oncogenic deregulation including some forms of enhancer hijacking. Across cancer genomes in the ICGC database[1] CTCF sites that form constitutive neighborhood boundaries were enriched for somatic mutations compared to other (non-boundary) CTCF sites (Hnisz et al. 2016). CTCF binding sites are frequently mutated in microsatellite-stable colorectal cancers with predicted consequences on CTCF binding affinity (Katainen et al. 2015). Observations on epigenetic deregulation provides additional evidence for a functional role of boundary element impairment in cancer: IDH mutant gliomas exhibit hyper-methylation of CTCF boundaries and this phenotype was found to be associated with upregulation of the PDGFRA oncogene, an effect that could also be provoked by targeted disruption of a specific boundary element at its locus (Flavahan et al. 2016). A pan-cancer study of somatic rearrangements across cancer types implicated structural variation involving TAD boundaries in the upregulation of IRS4 in several cancers and IGF2 in colorectal cancer (Weischenfeldt et al. 2017). Weischenfeldt et al. attributed upregulation of IGF2 to duplications forming a de-novo TAD that isolates IGF2 together with a distal super-enhancer element. Similarly, the authors found upregulation of IRS4 to be associated with a TAD boundary deletion, possibly fusing two adjacent chromatin domains and facilitating contacts between IRS4 and non-cognate enhancers of the neighboring domain. A computational model predicted CTCF driver mutations by their functional impact and recurrence in 1,962 whole genomes of 21 tumor types, identified 21 candidate insulator elements as targets for driver mutations and found such drivers to be associated with differential expression of TGFB1 across different cancers (E. M. Liu et al. 2019).

SCNAs coincide with SVs, as they are associated with structural changes. However, independent from gene regulatory effects that might occur due to structural rearrangements, as described above, SCNAs influence gene expression in a dosage-dependent manner. Here, the difference in the amount of gene copies available to the transcriptional machinery alters the resulting RNA levels in the cell. Copy-number gains and amplifications increase the amount of genetic material, while losses reduce the amount of DNA. Consequently losses of gene copies can lead to reduced gene expression, whereas additional copies increase expression levels, due to higher availability of DNA template material. Strong copy-number increases, known as amplifications of 10 or more copies, introduce extremely abundant DNA sequences of the same genomic region and can result in high expression levels of genes transcribed from these regions. SCNAs are prevalent in many cancers, are

[1] http://dcc.icgc.org/ (accessed 4 Nov 2020)

expected to be under selection in tumor evolution and to underlie pathogenic deregulation of transcriptomes. Comparison of SCNAs between epithelial neoplasms revealed frequent genomic imbalances due to arm-level copy number gains and losses, with both common and site-specific preferences (Baudis 2007), indicating that SCNA formation is a non-random process likely under selection during tumor growth. A global partitioning of genomic pan-cancer profiles discriminates between tumors driven by point mutations and SCNAs (Ciriello et al. 2013), showing that these two distinct groups result from different mutational processes fueling separate genetic paths in cancer pathogenesis. Across different cancer types typical somatic cancer genome contains 25% arm-levels SCNAs and 10% smaller, segmental or focal SCNAs (Beroukhim et al. 2010), so that often a substantial proportion of genes is expected to be affected by SCNAs in a given tumor genome. Analyses integrating copy-number variation and gene expression showed marked correlation between these measurements in breast cancer (Bergamaschi et al. 2006; Horlings et al. 2010) and gastric cancer (Junnila et al. 2010; L. Cheng et al. 2012; B. Fan et al. 2012). These studies used a correlation approach to identify genes under SCNA control and possibly differentially expressed between malignant and nonmalignant tissues due to this form of genetic regulation. Cancer-associated gene expression was found to be significantly associated with SCNAs across many cancers, showing that genetic dosage control is an important regulator of gene expression in tumor cells (Shao et al. 2019). After removing broad non-genetic components from over 77,000 expression profiles the residual expression in cancer samples was strongly driven by underlying copy-number differences, with almost all genes being sensitive to the copy-number effect to some degree (Fehrmann et al. 2015). Somatic gains generally increase gene expression and losses reduce gene expression (Shao et al. 2019). Shao and colleagues investigated differences between genes with and without strong sensitivity to somatic copy-number dosage and between SCNA up- and down-regulated genes across the cancers considered. They showed that genes without notable copy-number dosage effect to be enriched for pathways involved in basal cell function maintenance. In contrast, copy-number up- and down-regulated genes were significantly associated with pathways of energy metabolism (up), ubiquitin mediated proteolysis (down) and wnt signaling pathway (down), indicating that genetic control by SCNAs favors certain cellular processes that may be important to cancer biology.

Focal amplifications are small SCNAs with strong copy-number gains. These alterations can induce dosage-dependent increases of gene expression and frequently upregulate key oncogenic drivers in several cancer types. Amplifications have a much stronger effect on

gene expression in comparison to copy-number gains (Shao et al. 2019). Because the alterations are relatively small (usually up to 3 Mb), it is easier to pinpoint specific gene candidates compared to larger SCNAs. A subset of cancers show characteristic amplification targets that sometimes describe cancer subtypes. Highly recurrent amplifications of ERBB2/HER2 are found in 18-25% of invasive breast carcinomas, MYCN is amplified in 20–25% of neuroblastomas, BCL2 in 31% of diffuse large B cell lymphomas, MLL/KMT2A and ALL1 in 5–10% of Acute myeloid leukemias, and FGF4 in 7% of gastric adenocarcinomas (Yi and Ju 2018). A pan-cancer analysis of amplification dependent overexpression identified amplifications in 40% of tumors and found recurrent amplifications of MYC and MET in 25% and 18% of colorectal cancers respectively, SKP2 in 21% of squamous cell carcinomas, HIST1H3B and MYCN in 19% and 13% of liver cancers, KIT in 57% of gastrointestinal stromal tumors and FOXL2 in 12% of squamous cell carcinomas (Ohshima et al. 2017). CCND1 is amplified in around 15% of breast cancers and recurrently involves amplifications of neighboring genes EMS1/CTTN and INT-2/FGF3 (Karlseder et al. 1994; Hui et al. 1997; Ormandy et al. 2003). These frequently amplified gene targets are involved in transcriptional regulation (ERBB2, MYC, MYCN, MLL/KMT2A, HIST1H3B, FOXL2), in anti-apoptotic or cell-cycle-promoting signaling (BCL2, CCND1, SKP2, KIT), belong growth factor families (ERBB2/HER2, FGF4, INT-2/FGF3) or transduce signals in established cancer pathways (ERBB2, MET, SKP2, KIT). Gene amplifications at lower frequencies are also prevalent across cancer types, associated with increased expression of target genes and enriched in kinases, cell cycle regulators, and MYC family members (Beroukhim et al. 2010) as well as epigenetic regulators (Zack et al. 2013). In a cohort of 2,197 breast cancers TP53 regulator MDM2, transcription factor MYC and growth factor receptor EGFR were found to be amplified in 5.7%, 5.3% and 0.8% of cases respectively (Al-Kuraya et al. 2004). Amplifications associated with unfavorable outcomes may also regulate genes involved in immune responses. For example the immunoglobulin receptor FCGR2B was found to be amplified in 3% of diffuse large B-cell lymphoma cases, and high expression of this gene was significantly associated with disease-specific survival and time to progression (Arthur et al. 2018). Taken together these findings suggest that gene amplifications are important somatic copy-number driver events found across different cancer types. Some cancers are characterized by amplifications of specific oncogenes. Gene amplifications with low recurrence are found across many cancer types and both high and low recurrence amplifications implicate regulatory factors or genes directly associated with cancer hallmarks, such as cell-cycle and growth promoting factors and regulation of genomic integrity by TP53.

In summary, different classes of non-coding somatic alterations can deregulate gene expression in cancer. Cancer genomes harbor a variety of somatic non-coding alterations, such as SNVs, SVs and SCNAs, of which a subset is expected to act as regulatory drivers. Such driver mutations cause changes in expression of genes associated with disease-related processes. Mechanisms by which these alterations deregulate genes include alterations of binding motifs of regulatory elements, changes to the gene's regulatory context, gene disruptions and copy-number dosage effects. SNVs can change transcription factor binding affinities, as shown for recurring point mutations of the TERT promoter, which are the most frequent non-coding SNV regulatory drivers across many cancer types known today. Somatic SV drivers translocate genetic elements to create new aberrant regulatory environments affecting transcription of target genes. Translocated elements include promoters and distal elements with epigenetic marks characteristic of potent enhancers. Driver mutations affecting chromatin structure, such as disruptions of insulators or SVs spanning TAD boundaries, put genes under control of aberrant chromatin domains, possibly involving non-cognate regulatory elements from neighboring domains. SCNAs deregulate expression of copy-number dosage-sensitive genes. Here, gains and losses typically lead to up- and downregulation of genes respectively. Focal amplifications are SCNAs with extreme copy-number gains and are known to induce strong upregulation of key oncogenic drivers.

Somatic alterations add an additional layer of genetic control to the underlying germline genome. As a result, genetic regulation of gene expression in cancer is the combined effect of the germline background and acquired somatic regulatory drivers. However, the interplay between germline and somatic regulation is not well understood. In embryonic tumors, such as neuroblastoma, somatic alterations are less likely to accumulate due to the patient's age. Here, germline regulatory variants may play an important role in predisposing to the malignancy. Additionally, non-coding somatic alterations that reoccur in tumors of the same type indicate that somatic deregulation is context specific. Expected differences in the genetic foundation between cancers show that it is important to consider regulatory effects across different classes of variation. Investigations of both germline and somatic components of regulation may help to gain a better understanding of etiology and underlying disease mechanisms.

2.3 Gene regulation in neuroblastoma

NB tumors display a remarkable sparsity of exonic SNVs, with very few genes recurrently affected by this class of somatic alterations (Pugh et al. 2013). However, NB tumors are frequently affected by SCNAs and SVs. These variants are less likely to change protein function, because generally, they do not introduce coding sequence changes. Still, they have the potential to modulate gene expression by alteratering regulatory elements, introducing copy-number dosage differences, or by disrupting transcription through gene truncation. Additionally, SNVs and other small alterations in the non-coding part of NB genomes remain largely unexplored, and their effect on gene regulation therefore unknown. The comparable low number of mutations affecting gene function indicate that neuroblastoma could be a disease mainly driven by deregulation of gene expression. Somatic aberrations that do not affect coding sequences may induce regulatory changes. Germline SNPs in non-coding regions were associated with NB susceptibility and high-risk tumors, indicating gene regulatory mechanisms to underlie disease predisposition and aggressiveness. Epistatic interactions between somatic and germline variants could form the basis for gene regulatory programs associated with neuroblastoma. Variation can directly affect gene expression through sequence-dependent modulation of regulatory element function, or indirectly, by modulation of epigenetic effects on gene expression. Lastly, epigenetic mechanisms could drive NB disease initiation and progression independent from variation through transgenerational inheritance of epigenetic patterns affecting gene regulation.

Somatic copy-number alterations in neuroblastoma modulate expression of affected genes by DNA-dosage. SCNAs are frequent somatic aberrations in neuroblastoma and distinct patterns of these alterations are observed in primary tumors (Brodeur, Sekhon, and Goldstein 1977; Brodeur et al. 1981; J. M. Maris et al. 1995, 2001). SCNAs comprise losses, gains and amplifications of genomic DNA. Copy-number alteration in neuroblastoma can affect whole chromosomes or chromosome arms, smaller genomic segments and focal loci, that often contain only one or a few genes (Janoueix-Lerosey et al. 2009; Squire et al. 1995). Early studies found focal amplifications of MYCN to be located on ecDNA (Kohl et al. 1983; Schwab et al. 1983), implicating DNA circularization as an important mechanism of genetic control. Other low frequency amplifications such as those of ALK and MDM2 were first identified in neuroblastoma cell lines (Corvi et al. 1995; Miyake et al. 2002), and evidence suggests that ecDNA may also play a role in their formation (Corvi et al. 1995). Simultaneous profiling of copy-number and gene expression in samples from multiple tumors

allowed to correlate segmental copy-numbers and expression of genes located on the altered genomic regions. Copy-number amplifications of MYCN and ALK were found to be associated with high expression levels of these oncogenes (Bordow et al. 1998; Y. Chen et al. 2008; Schulte et al. 2011; Qun Wang et al. 2006; Łastowska et al. 2007). Increased RNA levels were linked to copy-number gains at the LMO1 locus (K. Wang et al. 2011). Gains of chromosome arms 2p and 17q as well as losses at 1p, 3p, 4p, 10q and 11q were found to correlate with high and low expression of genes located on these chromosome arms respectively (Qun Wang et al. 2006; Łastowska et al. 2007). And gene expression in regions recurrently affected by such SCNAs were found to correlate with patient survival (Bordow et al. 1998; Łastowska et al. 2007; Schulte et al. 2011).

Chromosomal aberrations near driver genes can activate their expression by exposing the genes to a new regulatory context. Chromosomal instability in cancer cells causes DNA double strand breaks that are repaired in order to maintain the genomic integrity required for mitosis. DNA repair processes can join non-consecutive DNA molecules, thereby creating rearrangements both within and between chromosomes (Bunting and Nussenzweig 2013). Rearrangements close to the telomerase reverse transcriptase gene (TERT) in high-risk neuroblastomas frequently juxtapose the gene to a distal CRE, activating its expression (Peifer et al. 2015; Valentijn et al. 2015). Most of the re-arrangements involve enhancer elements identified by H3K27ac ChiP-seq; and many of these elements were compatible with the definition of super-enhancers (Chipumuro et al. 2014; Peifer et al. 2015; Valentijn et al. 2015). In the majority of cases identified in these studies breakpoints were located upstream of TERT close to its promoter region. However, Valentijn and colleagues also found TERT downstream rearrangements to translocate super-enhancers, suggesting that de-novo cis-regulatory interactions were not constrained to the promoter-proximal region. Similarly, expression of the MYCN homologue MYC was found to be controlled by trans-located or amplified enhancers in neuroblastoma cell lines and at low frequency (<2%) in patient-derived tumor samples (Zimmerman et al. 2018).

Structural variations disrupt expression of neuronal development genes in neuroblastoma. ATRX transcript structure was reported to be frequently affected by deletions and this alteration was associated with alternative lengthening of telomeres (ALT) (Cheung et al. 2012). Additionally, recurring SVs in ATRX, ODZ3 and PTPRD were identified at higher frequencies than expected by chance and these alterations were associated with reduced expression of affected genes (Molenaar, Koster, et al. 2012). Molenaar et al. discovered an

enrichment of GTPase-regulation in disrupted genes and found that this regulation specifically activated Rho or inactivated Rac, two GTPase families with opposing signaling in neuronal development. Rac signaling leads to axon extension and guidance, whereas Rho to collapse of the neuronal growth cone (Leeuwen et al. 1997). The disrupting variants were predicted to be under positive selection and induce defects in neuritogenesis (Molenaar, Koster, et al. 2012). Interestingly this study implicated ATRX in neuronal development, suggesting that its role in neuroblastoma may not be limited to telomere maintenance.

Risk-associated SNPs in non-coding regions predispose to neuroblastoma by imposing cis-regulatory effects on gene expression. Genome-wide association studies identified common SNPs in non-coding regions of the genome to predispose to neuroblastoma (John M. Maris et al. 2008; Capasso et al. 2009; K. Wang et al. 2011; Diskin et al. 2012; McDaniel et al. 2017). Since none of the SNPs identified or those in strong linkage was predicted to have coding consequences, regulatory mechanisms may underlie these associations. Studies that integrated SNP genotypes with gene expression linked risk SNPs to expression traits of proximal genes, confirming a regulatory role of risk associated loci: Gene expression of LMO1, NBPF23, LIN28B, let-7, MLF1 and MMP20 were linked to genotypes of proximal risk SNPs (K. Wang et al. 2011; Russell et al. 2015; Diskin et al. 2009, 2012; McDaniel et al. 2017; Chang et al. 2017). Risk SNPs in CASC15 and BARD1 were associated with expression of specific transcript isoforms with predicted oncogenic potential (Bosse et al. 2012; Russell et al. 2015); and risk SNPs near CPZ were found to be associated with promoter CpG methylation (McDaniel et al. 2017), suggesting epigenetic gene regulation under the control of genetic risk variants. Among these observations Oldrige et al. provided the strongest evidence for a risk locus to be implicated in cis-regulation: After imputing genotypes at the previously identified 11p15.4/LMO1 risk locus the top associated SNP was found to be located in an intronic enhancer element; and the SNP's protective allele was predicted to disrupt a motif of the GATA transcription factor family, leading to lower LMO1 expression (D. A. Oldridge et al. 2015). Table 1 lists loci associated with neuroblastoma traits in association studies and their functional implications on genes in cis.

Locus	SNP	Assoc.	Gene	Function	References
6p22.3	rs6939340, rs4712656 **rs9295534**	Case	CASC14, CASC15, **CASC15-S**	Isoform expression*	(John M. Maris et al. 2008; Pandey et al. 2014; Russell et al. 2015; McDaniel et al. 2017)
2q35	rs6435862, rs3768716, rs7587476, rs58430496, **rs10498026**	High risk	BARD1, **BARD1β**	Isoform expression	(Capasso et al. 2009; Bosse et al. 2012; McDaniel et al. 2017)
11p15.4	rs110419, **rs2168101**	Case	**LMO1**	Expression	(K. Wang et al. 2011; D. A. Oldridge et al. 2015)
1q23.3	rs1027702	Low risk	DUSP12	-	(Nguyen et al. 2011)
5q11.2	rs2619046	Low risk	DDX4, IL31RA	-	(Nguyen et al. 2011)
11p11.2	rs11037575, rs10742682	Low risk	HSD17B12	-	(Nguyen et al. 2011; McDaniel et al. 2017)
1q21	rs17162082, **CNV**	Case	**NBPF23**	Expression	(Diskin et al. 2009)
6q16	rs4336470, rs72990858	Case	HACE1	-	(Diskin et al. 2012; McDaniel et al. 2017)
6q16	**rs17065417**	Case	**LIN28B, let-7**	Expression	(Diskin et al. 2012; McDaniel et al. 2017)
17p13.1	**rs78378222**[†], rs35850753[†]	Case	TP53	Transcription termination	(Diskin et al. 2014)
4p16	**rs3796727**	Case	CPZ	Methylation	(McDaniel et al. 2017)
3q25	**rs6441201**	Case	RSRC1 **MLF1**	Expression	(McDaniel et al. 2017)
11q22.2	rs10895322	11q deletion	MMP20	Expression	(Chang et al. 2017)

Table 1: Neuroblastoma risk associated loci and their functional implication for genes in cis. Bold fonts indicate genes and SNPs in functional relation when multiple candidates are listed. (*) function suggested, but not verified. ([†]) rare variation.

Analysis of epigenetic deregulation by DNA methylation revealed prognostic relevant markers and disease-associated pathways in neuroblastoma. Methylation of CpG islands regulates gene expression by silencing of regulatory regions, such as promoters and enhancers. Early studies implicated methylation-induced silencing of RASSF1A and CASP8 in neuroblastoma, pointing towards the epigenetic deregulation of apoptosis pathways (Teitz et al. 2000; Astuti et al. 2001; Decock et al. 2011). Methylations of CpG sites at the PCDHA and PCDHB gene families, RASSF1A, BLU, HLP and CYP26C1 characterize the neuroblastoma CpG island methylator phenotype (CIMP), that was found to be associated with MYCN amplifications and with poor survival in samples without MYCN amplification (Abe et al. 2005, 2007). Analyses of CpG sites revealed additional prognostic methylation markers in neuroblastoma: Methylation patterns at loci of genes FOLH1, MYOD1, THBS1, FOXP1, RB1 and TDGF-1 were associated with unfavorable outcome (Lau et al. 2012; Ackermann et al. 2014; Yáñez et al. 2015; Ram Kumar and Schor 2018). The neuroblastoma-implicated, neuronal development gene PHOX2B was found to harbor aberrant promoter methylation in 2 of 13 neuroblastoma cell lines and 2 of 18 primary tumors investigated and hypermethylation was linked to lower expression of PHOX2B in these samples (de Pontual et al. 2007). Genes with the highest number of hypermethylated CpG sites include Telomerase reverse transcriptase (TERT), PCDHGA4, DLX5, and DLX6-AS1 and sites with variable methylation were found to be hypermethylated in NB tumors of unfavorable disease stage (Olsson et al. 2016). Olsson and colleagues reported hypermethylation of genes implicated in cell-adhesion and neuronal development pathways. Loss of imprinting of H19 and IGF2 is frequently found in Wilms' tumor and embryonal rhabdomyosarcoma, but mono-allelic expression of these genes indicated that imprinting is generally preserved in neuroblastoma (Wada et al. 1995).

Epigenetic profiling of CREs in neuroblastoma characterized core transcription factors of neuroblastoma cell identity. Groups of highly expressed transcription factors maintain the cell's regulatory program. Core cell identity transcription factors bind their own enhancer elements and thereby establish self-regulatory feed-forward loops termed core regulatory circuits (Young 2011). It was suggested that core regulatory circuit components can be identified by integrating maps of large clusters of enhancer elements, so called "super-enhancers", with gene expression (Whyte et al. 2013; Saint-André et al. 2016). ChiP-seq of histone modification H3K27ac in neuroblastoma cell lines revealed two enhancer states associated with a noradrenergic- and a neural crest-like (mesenchymal) cell identity (Boeva et al. 2017; van Groningen et al. 2017). Core regulatory circuit analysis of

super-enhancers and gene expression identified transcription factors as core components; PHOX2B, HAND2 and GATA3 were found to be associated with the noradrenergic identity, and AP-1 class transcription factors (JUN, JUNB, FOSL1 or FOSL2) with the neural crest-like identify (Boeva et al. 2017). Intermediate mixtures of the two cell identities may be present in the same cancer cell population, as reported for the neuroblastoma cell line SK-N-SH (Boeva et al. 2017; van Groningen et al. 2017), and for tumors *in vivo* (van Groningen et al. 2017). The expression of genes associated with the enhancer state activity can shift upon re-programming by mesenchymal transcription factors (van Groningen et al. 2017) and drug treatment (Boeva et al. 2017), suggesting a plasticity of cell identity between the two states as a mechanism of treatment response.

The transcription factor PHOX2B is a core regulator of noradrenergic cell identity in neuroblastoma and itself under cis-regulatory control by its super-enhancer region. PHOX2B is specific to the peripheral autonomic nervous system and expressed in most neuroblastoma cell lines and tumors (Stutterheim et al. 2008; Bielle et al. 2012; Boeva et al. 2017). Both PHOX2B knockdown (Ke et al. 2015; Boeva et al. 2017) and over-expression (Raabe et al. 2008) was reported to suppress neuroblastoma cell proliferation, indicating that its expression might be tightly controlled at levels optimal to cellular self-renewal. Both conserved and non-conserved genomic elements upstream of PHOX2B were found to contribute to tissue specific expression (McGaughey et al. 2009), indicating that the gene is controlled by a complex regulatory landscape with elements of different degrees of conservation. Transcription factors of the noraderenic module PHOX2B, HAND2 and GATA3 show strong co-occupancy in PHOX2B's super-enhancer region (Boeva et al. 2017). Even though PHOX2B is expressed exclusively in neuroblastoma cells of the noradrenergic identity, its super-enhancer is still prominent in some cells of mesenchymal identity (van Groningen et al. 2017), suggesting that its cis-regulation by PHOX2B enhancer elements might be important in maintaining of cell plasticity.

In summary, these studies show that cis-regulation of gene expression in neuroblastoma is controlled by a complex interplay between genetic and epigenetic factors. These local control mechanisms impact global traits, such as NB disease development and progression. Epigenetic deregulation of genes by DNA methylation contributes to cancer hallmarks, such as evasion of apoptosis and telomere maintenance. And it affects pathways of the neuro-developmental origin of the disease, specific to neuroblastoma but distinct from other embryonal tumors. Cis-regulation by DNA methylation is associated with expression of

genes of known implication in neuroblastoma, and its analysis identified new genes, with regulatory patterns indicative for disease outcome. Enhancer elements control the expression of transcription factors of neuroblastoma core regulatory circruites in feed-forward loops, implicating cis-regulation by these elements as key drivers of tumor cell identity. Common variation predisposes to neuroblastoma and cis-regulation of gene expression may be the underlying effect of risk-associated SNPs. Somatic SVs control neuroblastoma drivers by translocation of enhancer elements to proto-oncogenes and disruption of tumor suppressors. SCNAs impose major dosage-dependent effects on gene expression and these frequent somatic variants significantly contribute to disease progression. Focal amplifications upregulate key oncogenic drivers by strong DNA dosage increases. Amplifications of MYCN are frequently found on circular DNAs, implicating these extrachromosomal structures as important cis-regulatory drivers.

2.4 Telomere maintenance in neuroblastoma

Telomeres are regions of repetitive DNA sequences that protect the ends of linear chromosomes. In every cell cycle telomeres shorten. To prevent cellular senescence that occurs when telomeres become too short, proliferating cells must maintain these structures by elongation. Cancer cells are characterized by replicative immortality and therefore must have acquired a mechanism to maintain telomeres (Douglas Hanahan and Weinberg 2011). The enzyme Telomerase lengthens telomere ends by addition of telomere repeat sequences and TERT (Telomerase reverse transcriptase) is the Telomerase subunit that catalyses this process by reverse transcription. TERT is expressed in certain stem cells but deactivated in most differentiated cells (W. E. Wright et al. 1996). In cancer, activation of TERT is a common mechanism of telomere maintenance (N. W. Kim et al. 1994; Hahn et al. 1999). In many cancer types TERT is activated by somatic alterations, a mechanism that was first described in melanomas, where more than 70% of cases harbor two distinct promoter mutations that lead to activation of TERT (Franklin W. Huang et al. 2013; F. W. Huang et al. 2015; Barthel et al. 2017). In neuroblastoma TERT is activated by at least two distinct mechanisms, that are found in high risk tumors: The first mechanism is mediated by MYCN amplification, as tumors with this molecular phenotype show increased levels of TERT expression (Peifer et al. 2015; Hertwig, Peifer, and Fischer 2016), indicating MYCN-induced upregulated of TERT by transcriptional reprogramming. The second mechanism, described in section 2.3, involves somatic rearrangements that upregulate TERT expression in cis by hijacking of enhancer elements to the TERT locus (Peifer et al. 2015; Valentijn et al. 2015).

Alternative lengthening of telomeres (ALT) is a second maintenance mechanism that is found in many tumors that lack activation of TERT (T. M. Bryan et al. 1995; Tracy M. Bryan et al. 1997). ALT is based on recombination induced by breaks at telomeric DNA sequences (Dilley et al. 2016). In cancer, ALT is characterized by an excess of telomere repeats $(TTAGGG)_n$, extrachromosomal telomeric repeat sequences (Henson et al. 2009) and loss of function mutation in ATRX and DAXX genes (Heaphy et al. 2011; Sieverling et al. 2020). In neuroblastoma the ALT phenotype is significantly enriched in relapse cases and associated with poor outcome independent of the risk group, indicating that risk stratification could benefit from assessing this phenotype (Ackermann et al. 2018; Hartlieb et al. 2021). ATRX was found to be mutated in approximately 25% of high-risk neuroblastoma tumors by somatic deletion of exons 5-10 or somatic SNVs that introduce missense or nonsense mutations (Cheung et al. 2012; Koneru et al. 2020). ATRX alterations are associated with ALT and increased telomere length (Cheung et al. 2012; Valentijn et al. 2015; Peifer et al. 2015). 50-60% of ALT tumors harbored ATRX alterations (Koneru et al. 2020; Hartlieb et al. 2021) and both ATRX altered and wildtype ALT tumors showed reduced abundance of ATRX and DAXX (Hartlieb et al. 2021).

At least three mechanisms were described, by which loss of ATRX function could affect the ALT phenotype (S. L. George et al. 2020): ATRX deposits H3.3 histones at telomeres (Goldberg et al. 2010; Lewis et al. 2010) and thereby stabilizes chromatin by preventing the formation of DNA secondary structures, such as G-quadruplexes (Clynes, Higgs, and Gibbons 2013). Loss of ATRX function may therefore lead to the formation of DNA secondary structure and stalled replication forks, which are resolved by break induced recombination processes that facilitate ALT (Clynes et al. 2015). Additionally, altered ATRX interactions with the MRN complex, which resolves stalled replication forks by repair of double strand breaks, could modulate ALT activity (Clynes et al. 2015). A third mechanism involves the telomeric repeat-containing RNA (TERRA). In the absence of ATRX the telomeric TERRA is upregulated (Goldberg et al. 2010) and TERRA-induced formation of steady R-loops (DNA-RNA hybrids) between TERRA and telomeric DNA may induce a DNA damage, which is then repaired by homologous repair processes, that are associated with ALT (Graf et al. 2017). Figure 4 depicts the role of ATRX loss and in ALT.

In summary, telomere maintenance in neuroblastoma and other types of cancer is associated with activation of telomerase by either upregulation of TERT or induction of ALT. While telomerase maintains telomere length by reverse transcription, ALT lengthens

telomeres by break-induced homologous repair processes. In neuroblastoma ALT is associated with alterations of the ATRX gene. Compared to telomerase-dependent maintenance, ALT induces longer telomeres, detectable by an excess of telomere repeats.

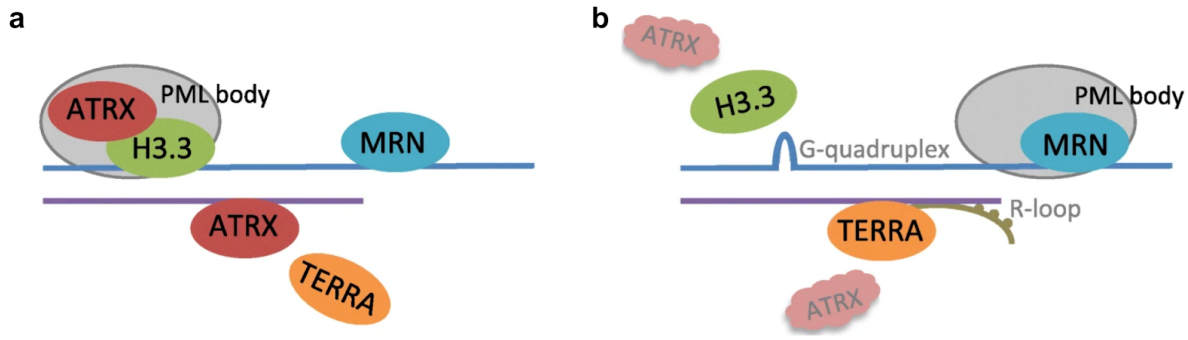


Figure 4: The role of ATRX loss in alternative lengthening of telomeres. **a**, Scheme of a normal telomere with colocalization of ATRX and H3.3 histones within PML bodies. **b**, In alternative lengthening of telomeres loss of ATRX function inhibits H3.3 deposition at telomeres and results in G-quadruplex formation; MRN co-localize with PML bodies and TERRA facilitates R-loop formation promoting DNA damage and homologous repair. Modified from original source: George, S.L., Parmar, V., Lorenzi, F. et al. Novel therapeutic strategies targeting telomere maintenance mechanisms in high-risk neuroblastoma. J Exp Clin Cancer Res 39, 78 (2020). This content is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

2.5 Extrachromosomal circular DNA

Extrachromosomal circular DNAs (eccDNA or ecDNA) were described in plants (Kinoshita et al. 1985; Cuzzoni et al. 1990; Cohen, Houben, and Segal 2007), yeast (Horowitz and Haber 1985; Sinclair and Guarente 1997; Henrik D. Møller et al. 2015), mammalian cells (Yamagishi et al. 1983; Kunisada et al. 1985; Flores, Moore, and Gaubatz 1987; Gaubatz and Flores 1990; Henrik Devitt Møller et al. 2018) and human cell lines (Radloff, Bauer, and Vinograd 1967; Kunisada and Yamagishi 1984; Cohen et al. 2010) as kilobase-sized small circular molecules, often termed eccDNA. Larger circular DNA (often termed ecDNA) were identified in tumors (D. Cox, Yuncken, and Spriggs 1965; Rausch et al. 2012; Turner et al. 2017), where they often occur in the form of double minute chromosomes (or double minutes) that are microscopically visible after staining metaphase DNA. Double minutes were first identified in neuroblastoma (D. Cox, Yuncken, and Spriggs 1965), and later associated with amplifications of drug resistance genes in murine cell lines (Kaufman, Brown, and Schimke 1979; Brown, Beverley, and Schimke 1981). Neuroblastoma cells contain different sizes of circular DNAs and these molecules are associated with oncogenic

amplifications in neuroblastoma and other cancer entities (Kohl et al. 1983; VanDevanter et al. 1990; Schwab et al. 1983; Rausch et al. 2012; Turner et al. 2017).

Different sources of circular DNA formation were suggested. In the episome model circular DNA is produced by excision of chromosomal DNA (Carroll et al. 1988; Storlazzi et al. 2006; Shibata et al. 2012). A concurrent theory proposed in that circular DNA arises from reverse transcription of RNA (Krolewski and Rush 1984). Cellular conditions or mechanisms that were predicted to promote circular DNA formation include DNA replication error (Schwab and Amler 1990; Paulsen et al. 2018), DNA damage (Rausch et al. 2012; Ly and Cleveland 2017; Henrik Devitt Møller et al. 2018; Verhaak, Bafna, and Mischel 2019) and transcriptional activity (Dillon et al. 2015). At least some circular DNAs contain repetitive sequences, which lead to the hypothesis that these molecules originate from sequence homology-based recombination (Jones and Potter 1985; Kunisada and Yamagishi 1987; Okumura, Kiyama, and Oishi 1987). Recent findings show that formation of ecDNA in cancer cells is associated with double strand breaks and subsequent DNA repair by non-homologous end joining (NEHJ) or microhomology mediated end joining (MMEJ) (Paulsen et al. 2020; Shoshani et al. 2020)[2]. The detailed investigation of structural variation in gene amplifications and mitotic segregation defects by Shoshani et al. showed that ecDNAs were created by NHEJ repair of shattered chromosomal fragments (chromothripsis).

Double minute chromosomes were associated with resistance to methotrexate in murine cell lines and the number of double minutes correlated with copies of the drug resistance gene (DHFR) (Kaufman, Brown, and Schimke 1979; Brown, Beverley, and Schimke 1981). Amplifications, as those introduced by double minutes, upregulate gene expression by an increase of gene dosage (W. H. Lee, Murphree, and Benedict 1984; Libermann et al. 1985; Nau et al. 1986; Wong et al. 1986). In a subset of neuroblastoma tumors the MYCN proto-oncogene is frequently amplified by double minute chromosomes (Kohl et al. 1983; Schwab et al. 1983), and these tumors often relapse after treatment and are associated with high mortality (Brodeur et al. 1984; Seeger et al. 1985; Bagatell et al. 2009). Analysis of copy-number and ASE in ecDNAs in a human glioblastoma cell line showed that high expression of associated oncogenes originated from the amplified allele (S. Wu et al. 2019). ecDNAs were found across a wide spectrum of cancer types and are the most common mechanism of gene amplification for many established oncogenes (Turner et al. 2017).

[2] Paulsen et al. 2020 refers to a preprint article.

Oncogenic amplification on ecDNAs facilitates tumor evolution by increased genetic heterogeneity, which can result in a growth advantage compared to amplifications that reside on chromosomes (Turner et al. 2017; deCarvalho et al. 2018).

Homogeneously staining regions (HSR) are stretches of identical giemsa staining visible on metaphase chromosomes. Mutually exclusive occurrence of HSRs and double minutes and identical staining properties in neuroblastoma cell lines lead to the conclusion that these two phenomena had a common origin (Balaban-Malenbaum and Gilbert 1977). Similar to double minutes HSR were found to carry amplifications of the drug resistance gene DHFR in a methotrexate-resistant cell line (Nunberg et al. 1978). Further studies suggested that double minutes can integrate into chromosomes (Schimke et al. 1978), which was later confirmed by time-series FISH experiments (Ruiz and Wahl 1990). Ruiz and Wahl also found that double minutes that integrated into telomeric regions destabilized affected chromosomes. Conversely, double strand breaks introduced within HSRs lead to the formation of double minutes (Coquelle et al. 2002). Finally, chromosomal integration in HSR and formation of new ecDNAs from fragments of destabilized, dicentric chromosomes in cell division was described in cancer cells in great detail (Shoshani et al. 2020). Similar to double minutes, oncogene amplifications by HSRs are common across many cancer types (Turner et al. 2017).

Taken together these sources show that circular DNAs are common in many organisms. In cancer, amplification-associated ecDNAs are formed by repair of broken chromosomal DNA fragments, that can result from segregation defects of dicentric chromosomes. ecDNAs together with HSRs, their intrachromosomal counterpart, are vehicles of oncogene amplification in neuroblastoma and other cancer entities. ecDNAs can integrate into chromosomes forming HSRs. Conversely, double strand breaks in HSRs and segregation defects of HSR-destabilized chromosomes can create ecDNAs.

2.6 Methodology

2.6.1 Next-generation sequencing

Since Frederick Sanger introduced a DNA sequencing method based on in-vitro replication of single DNA molecules (first-generation sequencing) a variety of high throughput methods were developed in the past two decades (Metzker 2010; Goodwin, McPherson, and McCombie 2016). These next-generation sequencing (NGS) methods have in common that

they can assay thousands of DNA sequences in parallel. The completion of the human genome sequence in 2003 delivered a comprehensive map for genetic studies in our species (International Human Genome Sequencing Consortium 2004). Tremendous efforts in annotating genetic elements in the human genome, such as genes and regulatory sequences, enabled the study of sequence variation, gene expression and gene regulation in health and disease at genome-wide scale (Curwen et al. 2004; Pruitt, Tatusova, and Maglott 2005; Harrow et al. 2006; ENCODE Project Consortium 2011). Since NGS is accessible to many researchers a plethora of assays have been developed that make use of the high throughput quantification abilities of this technology. Taken together, these developments introduced a new era to the research fields of molecular biology and bioinformatics. And in most research contexts, sequencing methods have now replaced DNA microarrays-based methods.

The sequencing data used in this work was generated by the Illumina/Solexa NGS technology[3] and I will therefore briefly describe this approach to DNA sequencing. In Illumina/Solexa NGS technology DNA fragments are first amplified and then sequenced in cycles of single nucleotide polymerizations, which results in short reads between 50-300 bp (Voelkerding, Dames, and Durtschi 2009). In a first step DNA fragments from a library preparation are linked to adaptor sequences. These adaptors can hybridize to oligonucleotides on a surface of the flow cell, the reactive chamber of the sequencing instrument. The flow cell contains millions of attached oligonucleotides anchors complementary to either of the two sequencing adaptor types linked to each of the DNA fragment's ends. The DNA fragment binds to these anchors and is amplified in a process termed bridge amplification. In this process the un-attached end of the DNA fragments bends over to a neighboring oligo, to which it is complementary, forming a "bridge". Now the hybridized second oligo acts as a primer for the replication of the reverse strand. After the double-stranded bridge is denatured the initial DNA fragment corresponds to two complementary sequences (forward and reverse strands) attached to the flow cell's surface. The process is repeated over and over and finally the reverse strands are cleaved, leaving clusters of forward strand sequences attached. In another process called sequencing-by-synthesis fluorophore-labeled nucleotide triphosphates are incorporated in the replication of the attached fragments. Only a single nucleotide is incorporated before the fluorophore is excited by a light source. Because the fluorophore wavelength is specific to the incorporated nucleotide's base, all synthesized sequences in a cluster emit the same

[3] www.illumina.com/technology/next-generation-sequencing.html (accessed 6 Aug 2020)

wavelength. The wavelength of the clusters are recorded and the fluorophore is cleaved, thereby completing the first cycle of the sequencing-by-synthesis process. In the subsequent cycle the next nucleotide will be incorporated. The emitted light of hundreds of million of clusters in the flow cell is recorded simultaneously, allowing for very high sequencing throughput. The length of the final sequencing reads are determined by the number of cycles, but this number is limited due to increasing measurement noise. Typically 50 to 300 cycles are run on modern Illumina sequencers, producing sequences of 50 to 300 bp each. Figure 5 shows hybridization, bridge amplification and sequencing-by-synthesis steps in the Illumina/Solexa NGS workflow. In short read sequencing of longer DNA fragments, paired-end protocols were developed to improve downstream analysis (Fullwood et al. 2009). In paired-end protocols sequence information is obtained from both ends of the DNA fragment, establishing the linkage of sequences at distances that exceed the sequence technologies read length limitations. In the paired-end workflow of Illumina/Solexa NGS the reverse strand is synthesized, amplified and sequenced after completion of the forward strand cycles. Thereby the sequence is additionally captured starting from the opposite end of the DNA fragment. Indices in the adapter sequences match resulting reads from both ends of the same fragment.

The procedure described above starts from a sequence library of DNA fragments, that can e.g. be prepared from genomic DNA of cell culture or tissue samples. In Whole Exome Sequencing (WES) fragments are enriched for DNA sequences of genes by specifically capturing and amplifying exonic fragments. WES produces high coverage (typically 100x) of exonic regions and is therefore suitable to detect e.g. rare variants and subclonal gene mutations in cancer. In Whole Genome Sequencing (WGS) the genomic DNA fragments are not enriched for specific regions and thus resulting reads cover both coding and non-coding regions. Due to the large non-coding proportion of the human genome even a moderate coverage requires a large number of sequencing reads (e.g. 30x requires around 600 million reads of 150 bp length). WGS is suitable to detect germline and clonal somatic variants in both genic and intergenic regions.

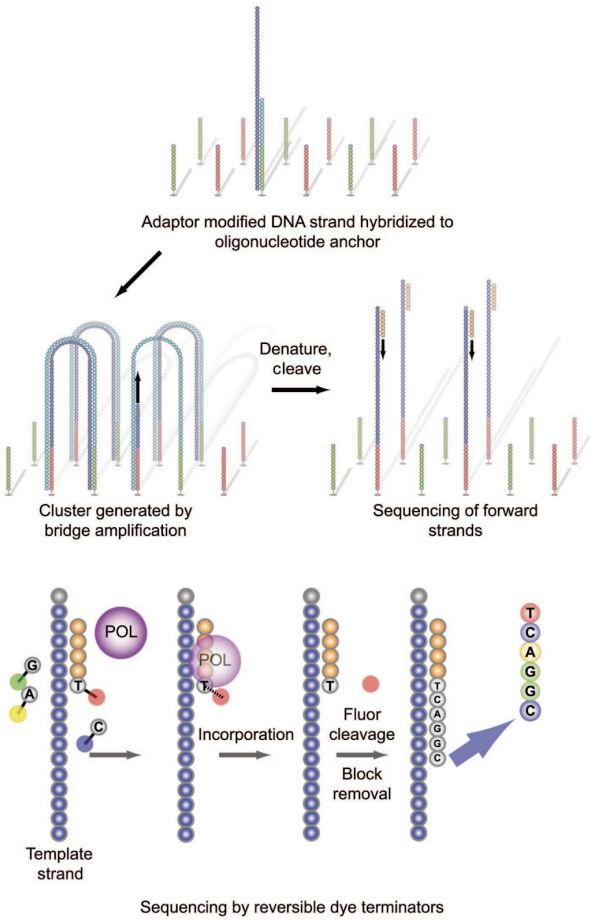


Figure 5: Overview of Illumina/Solexa DNA sequencing technology. Scheme depicts hybridization and bridge amplification followed by sequencing-by-synthesis of the forward strand. In each cycle one of four different fluorophore-labeled nucleotides is incorporated by DNA-polymerase (POL), detected and the fluorophore is cleaved before the next cycle begins. Source: Voelkerding, Karl V; Dames, Shale A, Next-Generation Sequencing: From Basic Research to Diagnostics, Clinical Chemistry, 2009, 55(4):641-58, by permission of Oxford University Press.

In order to analyze DNA sequencing reads in the context of their genomic origin, they need to be mapped to a corresponding reference sequence (or simply “reference”). Read mapping is the process in which each read or read pair is aligned to the reference. The resulting alignment contains mapping coordinates, mapping quality scores, and the original sequence and quality information per read. Several algorithms have been developed to efficiently align the vast number of short DNA reads resulting from NGS experiments (H. Li and Durbin 2009; Langmead et al. 2009). The human genome reference reflects the DNA sequence of chromosomes and the circular mitochondrial genome found in human cells as well as un-assembled contigs of sequences that could not (yet) be assigned to chromosomes. It represents a consensus “physiological” genome sequence that was derived from a few human individuals (*Nature Methods* 2010). Except for sequences in repetitive regions that

are undefined, the reference is unambiguous in that each position is represented by one of the four nucleotides. Therefore, it does not reflect the vast amount of variation that is found between individuals in the human population. Instead, human genome variants are defined as deviations from the reference sequence and stored in dedicated databases (Sherry et al. 2001; Bamford et al. 2004). Genomic instability in cancer is responsible for many somatic mutations, such as SNVs, copy-number changes and structural rearrangements. Rearranged genomes do not maintain the order of DNA sequences found in the reference. Due to inversions, deletions and translocations a given tumor DNA fragment can be composed of sequences that are non-contiguous in the corresponding germline genome. Somatic alterations pose difficulties in mapping reads from tumor DNA and RNA. Here, personalized references could potentially improve read mapping outcome. For example, de-novo assembled transcriptome references were previously used to detect gene-fusions and chimeric transcripts from cancer RNA (Mittal and McDonald 2017; Attig et al. 2019). However, constructing personalized genomes requires sequence assembly from whole genomic DNA at high coverage, which is still costly to obtain and computationally intensive. Furthermore, in cancer genomics, samples from different tumors are often jointly analyzed. E.g. frequencies of somatic or germline variants in a set of tumor genomes are subject to investigations. These comparisons require a common genomic coordinate system, but tumor genomes are highly heterogeneous, even between donors of the same type of cancer. For these reasons, the current standard procedure is to map tumor DNA and RNA reads to the human reference genome (He et al. 2020; PCAWG Transcriptome Core Group et al. 2020). Infact, many tools that identify structural variation are built to infer variation by artifacts, such as changes in read depth variance, discordantly aligned read pairs and split reads, that occur when rearranged DNA reads are aligned to the human reference sequence (Cameron, Di Stefano, and Papenfuss 2019).

To obtain sequence information from the transcriptome, RNA molecules are typically first transcribed to complementary DNA (cDNA) by the enzyme Reverse transcriptase. Reads from cDNA libraries can then be obtained by DNA sequencing on NGS platforms. This workflow is referred to as RNA sequencing (RNA-seq). Poly-dT primers or reagents to deplete ribosomal RNA are used to enrich mRNAs and non-ribosomal RNAs in RNA-seq library preparations respectively. Because of RNA splicing the cDNAs do not necessarily resemble contiguous genomic sequences, but may skip intronic regions. Therefore, either specific transcriptome references are required, or reads must be mapped to a genomic reference allowing larger regions in the reference to be skipped in the read alignment.

Specialized read mappers were designed to efficiently align RNA-seq reads from spliced RNA to the genomic reference with or without guidance by gene annotations (Trapnell, Pachter, and Salzberg 2009; T. D. Wu and Nacu 2010; Dobin et al. 2013; D. Kim, Langmead, and Salzberg 2015).

Due to the high throughput of NGS, resulting reads can be used for genome-wide quantifications. Here, reads aligned to genomic regions are counted and define the "coverage" of this region for the sequencing experiment. Read coverage may reveal enrichment of reads in specific genomic regions. Gene annotations, such as those from Ensembl and GENCODE define boundaries of genetic elements and the human gene annotations are regularly updated (A. D. Yates et al. 2020; Frankish et al. 2019). In RNA-seq-based quantification of gene expression the abundance of a gene's cDNA is determined by the read coverage in annotated exons of that gene. RNA-seq has by now replaced cDNA microarrays for the purpose of gene expression quantification in most research setups. However, particular care must be taken when comparing RNA-seq gene quantifications, because expression measures from different annotations are generally not comparable. Quantifications of mapped DNA from WGS or WES reads can be used to detect copy-number variation. And counts of DNA reads with sequence deviations at specific locations are used to calculate allele frequencies, which are the basis for calling SNVs, such as SNPs and somatic point mutations (somatic SNVs). Besides its application in the discovery of genetic variation, DNA sequencing by NGS has become the basis for many specialized protocols to measure a variety of phenotypes in context of the genomic sequence. The underlying principle of these protocols is the enrichment of particular DNA fragments in the sequencing library. I will here briefly describe enrichment strategies of three specialized DNA sequencing assays, that were used to generate data analysed in this thesis: ChIP-seq, ATAC-seq, and Circle-seq.

ChIP-seq. The chromatin immunoprecipitation (ChIP)-seq protocol enriches DNA fragments bound by specific proteins (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007). It is used to functionally characterize DNA sequences based on occupancy of the protein target. In a first step, proteins are cross-linked to the DNA sequence they occupy. The DNA is then sheared by sonication, producing DNA fragments with and without the linked protein under investigation. An antibody against the protein is used to specifically capture those protein-DNA complexes linked to the protein of interest in a process termed immuno-precipitation. The captured DNA fragments are then purified to

construct an NGS sequencing library. Mapped ChIP-seq reads reflect the genomic occupancy of the targeted protein. ChIP-seq is broadly applied to study epigenetic phenomena, such as occupancy of transcription factors and post-translational histone modifications (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007; ENCODE Project Consortium 2011).

ATAC-seq. Assay for transposase-accessible chromatin (ATAC)-seq can be used to measure the accessibility of DNA sequences with few input material in a relatively short time frame on a genome-wide scale (Buenrostro et al. 2013). Here, the mutated hyperactive transposase Tn5 is used to simultaneously create double strand breaks and ligate sequencing primers to DNAs in a process termed “tagmentation”, that was earlier described for WGS library preparation (Adey et al. 2010). Buenrostro and colleagues adapted the protocol to enrich tagmentation in accessible chromatin regions, by preserving protein-DNA interactions before Tn5 treatment. The coverage of mapped reads (and frequency of cut site coordinates) reflect how accessible a genomic region is to the transposase. It produces comparable enrichments to DNAse-seq, FAIR-seq and MNase-seq, which are earlier protocols for assessment of chromatin accessibility based on DNAse I treatment, protein-DNA cross-linking and nuclease digestions, respectively (Song and Crawford 2010; Giresi et al. 2007; Schones et al. 2008). Compared to these methods ATAC-seq requires lower amounts of input material and library preparation is accomplished in a relatively shorter amount of time, still yielding high quality results (Buenrostro et al. 2013). Many transcription factors are limited to bind regulatory regions in accessible chromatin. For this reason chromatin accessibility assays help to prioritize and annotate regulatory elements, such as promoters and enhancers, in cell lines, as well as healthy and disease-associated tissues (Boyle et al. 2008; ENCODE Project Consortium 2011; Thurman et al. 2012; Corces et al. 2018).

Circle-seq. Circle-seq is a protocol to enrich circular DNA molecules, such as ecDNA and eccDNA (Henrik D. Møller et al. 2015; Henrik Devitt Møller 2020). Here, circular DNA is first enriched by column chromatography of cell lysates. Then, an exonuclease treatment digests remaining linear DNA molecules. Lastly, a rolling circle amplification further enriches circular DNAs before the NGS library is prepared. The coverage of mapped Circle-seq reads can be used to identify regions that gave rise to circular DNA molecules. It is a promising tool to study genomic amplifications by ecDNAs in eukaryotes and human somatic tissues, including tumor samples (Henrik D. Møller et al. 2015; Henrik Devitt Møller et al. 2018;

Koche et al. 2020). As Circle-seq detects a characteristic of DNA itself, it can be seen as both somatic genotype and cellular phenotype.

Counts of mapped reads resulting from experimental enrichments, such as the ones described above, can be used to quantify the phenotype in defined regions of the genome. Additionally, genomic regions harboring the phenotype can be identified de-novo by methods that test for statistical enrichment of mapped reads in a process termed “peak calling” (Feng, Liu, and Zhang 2011; Ibrahim, Lacadie, and Ohler 2015). These methods first establish a background model to describe read count quantities expected to be observed by chance and then report genomic regions which exceed the read counts expected under the hypothesis of the background model. The resulting coordinates (or peaks) can be used in the annotation of genetic elements (such as promoters and enhancers) and selectively study the sequences of those elements.

Name	**Library source**	**Enrichment target**	**Application**
WGS	DNA	(None)	Identification of sequence variation, copy-number quantification
RNA-seq	RNA	Poly-adenylated or non-rRNA transcripts	Quantification of gene expression (transcript/exon expression, identification of splice sites)
Circle-seq	DNA	Circular DNA	Identification of ecDNA and eccDNA
ATAC-seq	DNA	Accessible DNA	Identification of regulatory DNA elements
ChIP-seq	DNA	Protein-interacting DNA	Identification of histone modifications (identification of transcription factor binding sites)

Table 2: Next-generation sequencing-based assays and their applications in this thesis. Examples of additional applications that are not used in this work are given in brackets.

In summary, the availability of the human genome sequence and the high measurement throughput of NGS have made it possible to study genetic variation and sequence-associated phenotypes genome-wide. Mapping and counting of reads is a versatile strategy and basis for variant discovery and sequence-based phenotype quantification. WGS has enabled us to simultaneously investigate nucleotide- as well as structural- and copy-number variation in the coding and non-coding genome. Comprehensive annotations of human genes and the RNA-seq technology support precise

quantifications of gene expression. Sequencing-based assays with specialized enrichment strategies were developed to study cellular phenotypes in the genomic context. Peak-calling methods employ statistical analysis of mapped read counts to determine coordinates of read enrichment and guide the annotation of genetic elements in the non-coding genome. In the work conducted in this thesis the data of different NGS-based protocols were analysed. Table 2 gives an overview of these assays and their applications.

2.6.2 Allele-specific expression

A substantial proportion of differences in gene expression between individuals in the human population is caused by heredible factors (Schadt et al. 2003; Morley et al. 2004). Genetic variation is inherited separately through the two parental haplotypes. The degree of heterozygosity in individual genomes therefore reflects the population's genetic diversity. In the diploid genome both gene copies are potentially expressed. However, due to genetic variation, imprinting and NMD the maternal and paternal allele can be expressed at different levels, resulting in imbalances in expression from the two alleles (S. E. Castel et al. 2015).

Allele-specific expression (ASE) analysis is a method to quantify expression imbalances between alleles by integrating germline genotypes and gene expression (H. Yan et al. 2002; Ge et al. 2009). In this approach, first heterozygous variants, for example heterozygous SNPs (hetSNPs), of an individual are determined. Heterozygous variants in expressed genomic regions give rise to two distinct populations of RNA molecules. The difference in the RNA sequence between these populations reflects the RNA's allelic origin. ASE quantifies the amount of RNA molecules from the respective alleles by these sequence polymorphisms. If the haplotype (the sequence of genotypes on the same allele) is known, allelic information across multiple heterozygous variants for the same gene can be aggregated to improve accuracy of ASE quantification. ASE at hetSNPs can be determined by hybridization of RNA with SNPs arrays (Ge et al. 2009; Campino et al. 2008) or by RNA-seq experiments (Main et al. 2009; Degner et al. 2009). Under the assumption that the quantification of RNA of neighboring SNPs is technically independent, gene-level ASE can be determined by first calculating allelic counts at expressed hetSNPs and then summarizing allelic counts of hetSNPs overlapping the same gene.

The ASE ratio is a measure of the allelic expression imbalance measured at a hetSNP (Q. Li et al. 2013) and can be generalized to summarize multiple SNPs per gene:

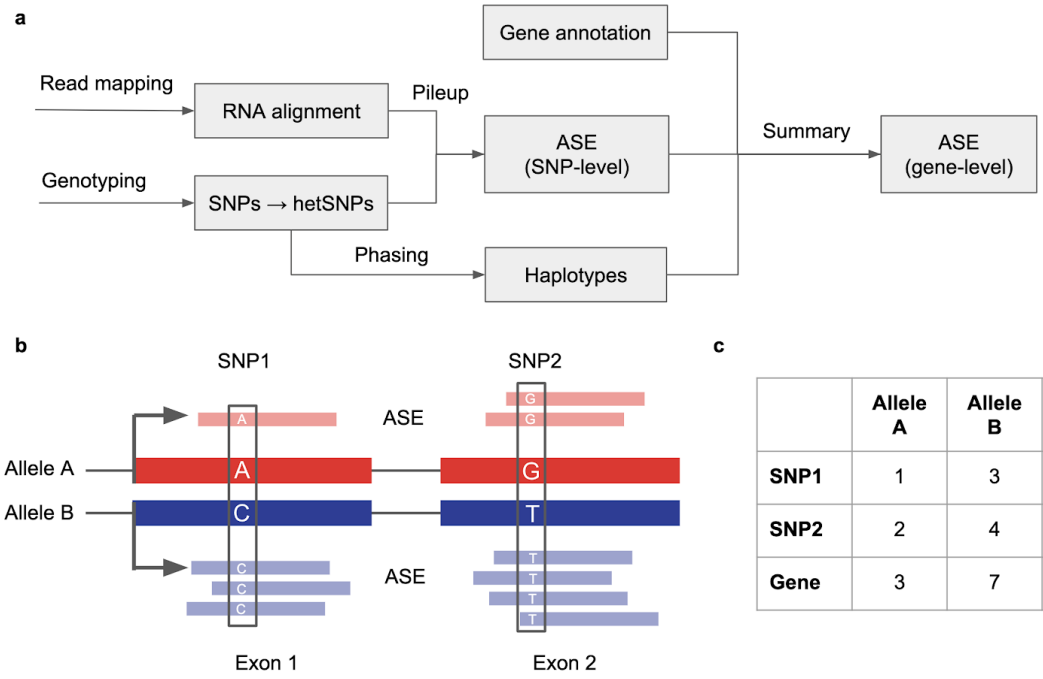
$$r_i = max(a_{i1}, a_{i2}) / (a_{i1} + a_{i2}) \quad (1)$$

$$\text{with } a_{ij} = \sum_k s_{kj} m_{ki},\ j \in \{1, 2\}$$

Here r_i is the ASE ratio for gene i, a_{ij} the aggregated allelic counts for gene i and allele j, s_{kj} the allelic count of allele j and SNP k and m_{ki} the entry in the $N_k \times N_i$ membership matrix M of SNPs and genes, where $m_{ki} := 1$ if SNP k overlaps with gene i and 0, if not not. Matrix M is determined by the gene annotation. Taking the maximum of the two allelic counts instead of picking one ensures that higher ASE ratios reflect stronger expression imbalances, regardless which allele is dominantly expressed. The allelic identity j across multiple allelic SNP counts s_{kj} requires the reconstruction of haplotypes, which can be achieved by statistical methods on population data (Stephens, Smith, and Donnelly 2001; Browning and Browning 2007). In RNA-seq-based quantification of ASE at hetSNPs mapped reads are counted at positions of hetSNPs of the two alleles separately in a process referred to as “pileup”. Figure 6 depicts an ASE workflow based on the pileup of RNA-seq reads at hetSNPs and gives an example on how SNP counts are summarized to a gene-level ASE phenotype for a gene overlapping two expressed hetSNPs.

Another workflow to determine gene-level ASE is based on the alignment of RNA-seq reads to two separate references representing the maternal and paternal transcriptome (Rozowsky et al. 2011; Turro et al. 2011). Reads that map with higher confidence to one of the two references are counted for the respective allele, other reads are discarded. Gene-level ASE can then simply be determined by counting the reads from the two separate alignments in the coordinates of the gene. In experiments where paternal and maternal chromosomes are known (e.g. in controlled laboratory experiments of gene expression in F1 individuals from two different inbred parent strains) the references can be constructed based on the homozygous variants of the F0 individuals. However, if the parental haplotypes are unknown, parental references can still be reconstructed by statistical methods. Even though this will likely not reconstruct the true haplotypes of whole chromosomes, it may still be accurate enough in the limits of most annotated genes. An advantage of methods that are based on alignments to two parental references with separate read counting is that they are less prone to mapping biases introduced by the reference and they resolve the problem of

double-counting reads overlapping multiple SNPs. A disadvantage is that they do not per se determine ASE at the SNP level and the construction and alignment to two parental references is computationally much more demanding.



	Allele A	Allele B	ASE ratio
SNP1	1	3	0.75
SNP2	2	4	0.66
Gene	3	7	0.70

Figure 6: Allele-specific expression is determined at expressed heterozygous SNPs and aggregated to gene-level results. **a**, A workflow to determine gene-level ASE from RNA-seq and genotypes by pileups of reads at expressed hetSNPs. **b-c**, Haplotype reconstruction (phasing) allows to determine gene-level ASE by summing allelic counts on the same allele. **c**, Allelic-counts and ASE ratios per SNP and at gene-level for the example given in (**b**).

ASE specifically captures cis-regulatory effects, because it controls for trans-regulation and technical variation between samples. Besides genetic variation, interactions between genes and environment as well as technical artifacts are additional sources of gene expression variation between samples. Environmental conditions can trigger regulatory programs that control expression of target genes in trans (López-Maury, Marguerat, and Bähler 2008). Technical artifacts introduce biases that do not reflect biological phenomena but sampling differences often seen as batch effects. A comparison of total gene expression phenotype between individuals is particularly prone to these sources of variation. In contrast, the ASE phenotype is controlled by the genomic locus and the cellular environment (Pastinen 2010). It is therefore less sensitive to environmental conditions and technical biases that exert their effects on both alleles simultaneously. This property makes ASE a promising tool to investigate cis-regulation of gene expression.

Cis-effects that induce ASE include regulatory polymorphisms, DNA methylation, copy-number imbalances, regulatory variation and variants leading to nonsense-mediated mRNA decay (NMD). The scheme in figure 7 depicts how copy-number imbalance and regulatory variation influence ASE. In functional genomics studies ASE was used to detect cis-effects induced by regulatory variation (McCarroll et al. 2008; Ge et al. 2009; Fogarty et al. 2010; Lappalainen et al. 2013; Battle et al. 2014). DNA methylation of intronic and promoter-proximal CpG sites can cause ASE of target genes (Milani et al. 2009). Parent-of-origin imprinting is a methylation-based mechanism of gene regulation inducing ASE by silencing expression of the maternal or paternal allele (Sakatani et al. 2001; Pollard et al. 2008). Genomic imbalances as a result of somatic losses and gains of copy-number segments introduce ASE by differences in allelic dosage (Tuch et al. 2010; PCAWG Transcriptome Core Group et al. 2020). Individual somatic aberrations, such as somatic SNVs, rearrangements and focal amplifications are expected to frequently affect single alleles. Consequently, those aberrations that affect gene expression lead to ASE of their target genes. Somatic SNVs in the promoter of TERT, common in many cancer entities, lead to mono-allelic expression of the TERT (F. W. Huang et al. 2015). MYCN amplifications were found to be frequently mono-allelic (J. M. Cheng et al. 1993) suggesting that MYCN is subject to strong ASE in tumors harboring focal amplifications of this oncogene. Expression imbalances of oncogenes activated by somatic rearrangements involving enhancer elements indicate that the activation is constrained to the rearranged allele (Northcott et al. 2014; Peifer et al. 2015; Gryder et al. 2020). NMD is a cellular surveillance mechanism that degrades mRNA transcripts harboring nonsense codons (Maquat 1995). NMD variants create nonsense codons that lead to mono-allelic transcript degradation and subsequently to ASE of the affected gene. ASE was used to identify widespread NMD associated with variation in human cell lines (Lappalainen et al. 2013; MacArthur et al. 2012) and cancer samples (Lindeboom, Supek, and Lehner 2016). These findings show that cis-regulation drives ASE of affected genes and suggest that this phenotyp is as a suitable tool to detect genetic and epigentic factors regulating gene expression.

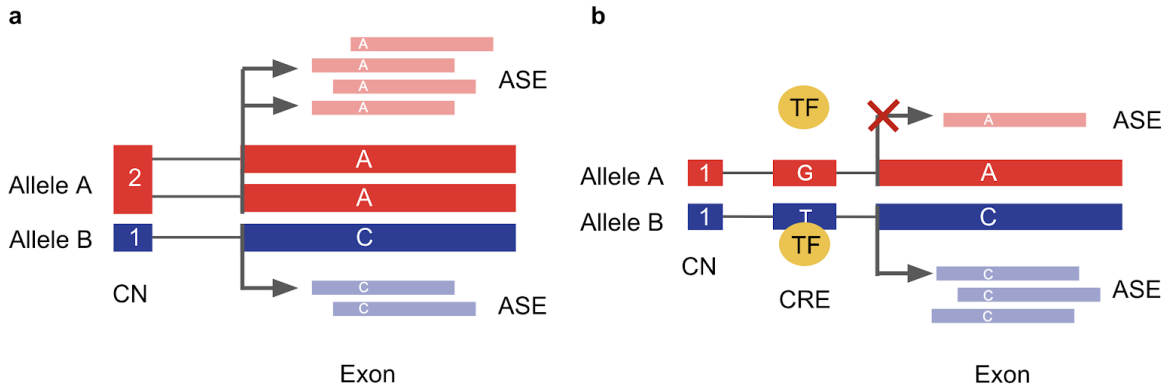


Figure 7: Allele-specific expression induced by copy-number imbalance and allelic differences in CRE activity. ASE induced by (**a**) copy-number imbalance and (**b**) modulation of a regulatory element through a single nucleotide variant. ASE is determined by the overlap of sequencing reads (light red and light blue) with an expressed heterozygous SNP genotyped as A/C. CN: copy-number, CRE: cis-regulatory element, TF: transcription factor.

ASE can detect cis-regulation of unknown origin. Specific genetic and epigenetic differences between alleles were found to be associated with ASE. These investigations helped to pinpoint cis-acting variation. The ASE phenotype reflects differences of gene expression in cis specifically because it is well controlled for bi-allelic including trans-regulation and technical and environmental variation between samples. If a gene is expressed differently between alleles, a cis-regulatory effect can be assumed even without identifying its source. ASE can therefore reveal cis-regulation of unknown origin by a “phenotype-first” approach. It has been applied in this way to identify cis-regulation undetectable by eQTL analysis (Ge et al. 2009) and to pinpoint cis-regulation in cancer (Milani et al. 2007; Ongen et al. 2014; Przytycki and Singh 2020).

Lack of informative genes and mapping biases limit the use of ASE. Heterozygous variants overlapping transcribed regions can be obtained by genotyping. ASE analysis requires reliable genotypes because false-positive heterozygous variants cause extreme outliers in allelic ratios of RNA: A true homozygous variant, falsely genotyped as heterozygous, results in an extreme ASE phenotype because all RNA is assigned to one of the alleles. Therefore ASE analysis requires stringent filtering of genotypes, which may in turn reduce the number of true positive heterozygous variants. If no expressed heterozygous variant is identified in a given sample, then the sample is not informative for ASE of that particular gene. These “sample dropouts” cause ASE to be a sparse measure. For a given gene, only a number of samples will be informative for ASE, namely those samples that have a least one expressed

heterozygous variant. This sparsity limits the statistical power in tests based on the ASE phenotype. In some cases, genes cannot be considered, because the number of informative samples drops below a threshold required for genome-wide testing. Additionally, statistical p-values between genes are difficult to compare, because of varying sample sizes per gene. Another limitation to ASE is its susceptibility to mapping biases. In short read mapping biases skew allelic read counts. Commonly used read mapping algorithms use a linear reference sequence per chromosome to represent the genome. Because in this representation the reference sequence does not reflect alternative variants the mapping procedure is inherently biased towards the reference allele (Degner et al. 2009). A variant read will contain at least one additional mismatch to the reference sequence, systematically reducing its mapping quality. In ASE, this so-called reference bias leads to the overestimation of expression from the reference allele on average but varies in effect and direction of bias depending on the sequence context. Strategies to counteract such biases include the consideration of SNPs during read mapping, filtering of unreliable SNPs, and the normalization of ASE ratios (Degner et al. 2009; Yuan and Qin 2012; Z. Liu et al. 2014). Sample dropouts and sequence biases may limit the application of ASE. In some instances, these obstacles may prevent cis-regulated genes to be detected by the ASE phenotype.

In conclusion, ASE is a well-controlled measure for cis-regulation. The two parental alleles of a gene can be expressed at different levels due to genetic and epigenetic factors. ASE quantifies these differences in expression between the alleles by integrating data of heterozygous variation (typically SNPs) with gene expression quantification. Haplotypes are required to make use of multiple heterozygous variants for gene-level ASE and, if unknown, can be constructed by statistical methods. ASE captures cis-effects on gene expression introduced by regulatory variants, methylation, and copy-number imbalances but also potentially unknown cis-effects. The lack of heterozygous variation and mapping biases limits the use of ASE, but the latter can be accounted for by filtering and normalization techniques.

2.6.3 Copy-number analysis in tumors

Eukaryotic cells contain pairs of homologous chromosomes (with the exception of gonosomes in males), one inherited from each parent. The state of chromosomes in a cell is referred to as the karyotype and the karyotype of a complete set of chromosomes is considered euploid. Aneuploidy describes a karyotype that differs in chromosomal copy-number from the euploid karyotype and in humans is often found in developmental

diseases and tumor cells. Aneuploidy encompasses gains and losses of whole chromosomes, chromosome arms and in a quantitative definition also smaller gains and losses (Ben-David and Amon 2019). In cancer aneuploidy is associated with chromosomal instability (CIN) a form of genomic instability (GIN) which facilitates errors in chromosomal segregation during mitosis, resulting in variable aneuploid karyotypes of daughter cells (Sansregret and Swanton 2017; Chunduri and Storchová 2019). CIN may originate from defects in mitotic checkpoints, microtubule attachment, mitotic spindle or chromosome cohesion and is associated with tumor progression, relapse and drug resistance (reviewed by Chundrui and Storchová). Aneuploidy in cancer is reflected by SCNAs of chromosomes and chromosome arms and these alterations may be important drivers of tumor evolution (Sansregret and Swanton 2017). It was suggested that aneuploidy itself may contribute to CIN, so that once SCNAs are acquired, they can promote additional SCNAs in subsequent cell divisions by gene-dosage induced perturbations of the mitotic protein machinery, forming a positive feedback loop (Potapova, Zhu, and Li 2013; Giam and Rancati 2015). Infact it has been shown that CIN contributes to intra-tumor heterogeneity (Navin et al. 2011; T.-M. Kim et al. 2015; de Bruin et al. 2014; L. R. Yates et al. 2015) and studies in yeast provided evidence for superior adaptability of aneuploid cells to stress conditions compared to their diploid counterparts (G. Chen et al. 2015; Selmecki et al. 2015), underlying the hypothesis, by which CIN generates karyotype diversity that increases the sampling space for evolutionary adaptation.

Qualitatively aneuploidy is distinguished from smaller gains and losses of copy-number. These smaller alterations comprise segmental and focal CNAs. Segmental CNAs are smaller than chromosome arms and often defined as larger than ~5 Mb. Focal alterations are usually smaller than 5 Mb. An example of a focal alteration in neuroblastoma are the amplification of MYCN and surrounding genes on ecDNAs and the smaller deletion affecting a set of ATRX exons (Section 2.3 and 2.5). The qualitative definition of aneuploidy implies that tumor genomes harboring one or multiple focal alterations are still considered euploid. However, quantitatively these smaller alterations still affect the ploidy. For example, in a euploid genome, that contains two copies of each of the homologous chromosomes, the quantitative ploidy value is equal to the average number of homologous copies (ploidy = 2n). A genome harboring focal amplifications but that is otherwise diploid will have an increased ploidy value (ploidy > 2n) depending on the copy-number of the amplifications. In cancer both aneuploidy and smaller alterations introduce SCNAs, that have implications on phenotype and prognosis and that are therefore extensively studied using different approaches.

Traditionally, chromosomal copy-number alterations were investigated by karyotyping methods that visualize mitotic chromosomes in light microscopy. Here, after cells have been arrested in metaphase, DNA is stained and karyotypes are inspected in light microscopy to investigate alterations and structural abnormalities. These methods make use of dyes that either stain nucleic acid unspecifically, such as the Giemsa stain, or sequence-specific by fluorescent in-situ hybridization (FISH). FISH utilizes fluorophore-labeled DNA probes to target specific sequences. In spectral karyotyping chromosomes are “colored” by FISH probes that carry combinations of fluorophores specific to the chromosomal origin of their target sequence (Schröck et al. 1996; Imataka and Arisaka 2012). Spectral karyotyping results in chromosome maps in which colors distinguish the chromosomal source of genetic material (see Schröck et al. figure 2 and 3 for examples of euploid and aneuploid karyotypes respectively). In contrast to the Giemsa staining method, SKY is particularly suitable to identify translocations of genetic material between chromosomes.

With the increased availability of sample material and the expanding research interest in comparative genomics, automated methods for quantification of copy-number variation were developed. Such methods measure the abundance of DNA at genomic loci and are able to detect even small copy-number variation, such as those introduced by focal gains and losses in the range of a few megabases. Techniques that make use of donor-matched reference and control samples are frequently used to study SCNAs and are of particular interest in cancer genomics. Here, the DNA of a test sample is derived from a tumor, and the reference sample from a normal tissue of the same donor. Comparative genomic hybridization (CGH) is a technique to measure copy-number variation by comparing FISH signals from two DNA samples (Tanner et al. 1996). In CGH the sample DNA (reference and test) is labeled by two different fluorescent dyes and denatured to generate probes. These probes are mixed and applied to a DNA template so that they compete for hybridization to their specific target sequence on the template. Abundances of DNA from reference and test samples are quantified by shifts in the fluorescent signal towards the respective dye’s wavelength. Originally, metaphase chromosomes were used as DNA templates for hybridization, which limited the resolution of CNV detection. Array CGH was developed to overcome these limitations. Here, probes are hybridized to short immobilized DNA sequences on glass slide matrices (arrays), which improves both resolution and the automated analysis of fluorescence signals (Solinas-Toldo et al. 1997; Pinkel et al. 1998). Similar to the comparison of hybridization signals, in NGS the differences in read depth at genomic positions between alignments of DNA reads from reference and test samples

reflect the relative abundance of DNA. In the last decade, numerous computational tools were developed that exploit read depth differences between samples in WES or WGS to determine genome-wide copy-number measurements (Sathirapongsasuti et al. 2011; Boeva et al. 2012; Klambauer et al. 2012; Koboldt et al. 2012; J. Li et al. 2012; Amarasinghe, Li, and Halgamuge 2013; Amarasinghe et al. 2014; Talevich et al. 2016; Kong et al. 2017).

Abundance measurements of genomic DNA overlapping heterozygous variation yield allelic copy-number information that complements abundance measurements in homozygous regions. In genomic regions with loss of heterozygosity (LOH) only one of the two original genotypes is observed. LOH cannot be determined by coverage in homozygous regions alone, because more than one copy of the retained allele might be present (copy-number neutral LOH). Similarly to the allelic preferences in regions of LOH imbalanced gains, losses and amplifications result in one of the parental alleles being over-represented. These allelic skews can be measured at heterozygous variants such as hetSNPs. High-density SNP arrays were originally designed for genotyping but have been used to determine allelic imbalances in order to investigate LOH and imbalanced CNAs in tumor samples (Lindblad-Toh et al. 2000; Bignell et al. 2004) and computational tools were developed to infer allele-specific copy-number (ASCN) profiles from SNP array signals (Popova et al. 2009; Van Loo et al. 2010; Rasmussen et al. 2011). Similarly, more recent methods utilize read coverage information at hetSNPs from NGS sequencing to estimate ASCN profiles (Mayrhofer, DiLorenzo, and Isaksson 2013; Favero et al. 2015; Raine et al. 2016; Shen and Seshan 2016).

Array- and sequencing-based methods allow high-throughput copy-number investigation across many samples. However, between-sample comparison of copy-number information is hampered by the fact that tumor samples show variable ploidy and are heterogeneous in their cell composition (Witz and Levy-Nissenbaum 2006; Chunduri and Storchová 2019). For example, differences in tumor cell content and normal cell admixture make it difficult to directly compare DNA abundance measures between such heterogeneous tumor samples. Additionally, hybridization signals and coverage depths are continuous measurements that do not directly translate into integer numbers of the two parental alleles. To infer such profiles in tumor samples, both tumor cell fraction (or purity) and ploidy values of the sample need to be determined. To reliably construct integer copy-number profiles methods such as ASCAT, FACETS and Sequenza estimate purity and ploidy values from the input data (Favero et al. 2015; Raine et al. 2016; Shen and Seshan 2016). After obtaining such

estimates, the allelic copy-number can be inferred. In the ASCAT algorithm the coverage difference between tumor and normal and the allelic skews are modeled by purity and ploidy estimates and the allelic copy-numbers at each hetSNP position i (Van Loo et al. 2010):

$$r_i = \gamma \log_2 \left(\frac{2(1-\rho) + \rho(n_{A,i}+n_{B,i})}{\Psi} \right) \quad (2)$$

$$b_i = \left(\frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho(n_{A,i}+n_{B,i})} \right) \quad (3)$$

Here, r_i is the measured ratio between the coverage of normal and tumor sample, b_i the measured B-allele frequency, ρ the estimated purity, Ψ the average sample ploidy given by $\Psi = 2(1-\rho) + \rho\Psi_t$, with Ψ_t the tumor ploidy, $n_{A,i}$ and $n_{B,i}$ the integer allelic copy-numbers for the A and B allele respectively, and γ a platform specific parameter. On the basis of equations 2 and 3 ASCAT determines estimates for the allele specific copy numbers. Figure 8 shows purity and ploidy estimates and corresponding ASCN profiles from an ASCAT analysis of two breast carcinoma samples based on measurements of an Illumina 109K SNP array.

ASCN analysis has major advantages over the characterization of total copy-number or classification of CNVs by broader classes such as gains, losses and amplifications. First, ASCN profiles reflect the copy numbers from both alleles, which allows for a comprehensive annotation of genomic imbalances from chromosome- to gene-level. Second, more specifically it allows the identification of LOH events including those that are copy-number neutral and therefore invisible to analyses based on total copy-number alone. And third, it can provide better estimates for tumor ploidy and purity, due to the complementary information provided by allelic skewness (Shen and Seshan 2016). In utilizing hetSNPs to characterize genomic imbalances ASCN- complements ASE analysis and may reveal copy-number driven expression imbalances. For these reasons it is an ideal tool to investigate the genomic-basis of ASE and distinguish copy-number dependent and independent ASE.

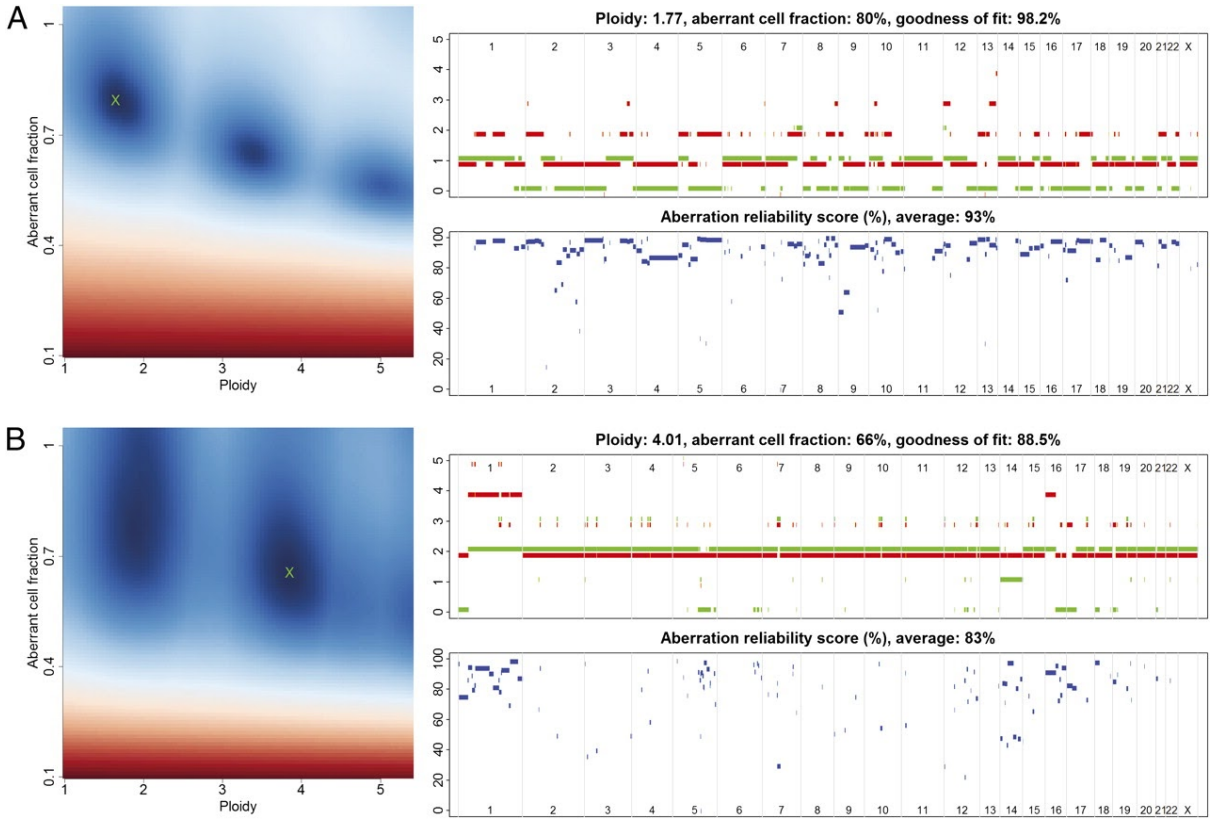


Figure 8: Tumor purity and ploidy estimates and corresponding allele-specific copy-number profiles obtained by the software ASCAT for two breast carcinoma samples based on Illumina 109K SNP array data. Results for (**a**) a tumor sample with estimated ploidy close to 2n and (**b**) a tumor sample with estimated ploidy close to 4n. Left: ASCAT determines goodness of fit (red low, blue high) for combinations of purity and ploidy values and selects the best fit (green cross). Right: The resulting profiles show copy-number gains and losses and the respective integer counts of major (red) and minor (green) alleles. For better readability allelic counts are slightly shifted up and down relative to their actual integer value respectively. An aberration reliability score is calculated for aberrations that deviate from the estimated ploidy. Source: Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., & Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America, 107(39), 16910–16915. https://doi.org/10.1073/pnas.1009843107. By permission of Proceedings of the National Academy of Sciences of the United States of America.

In conclusion, SCNAs in tumors originate from CIN-induced aneuploidy and smaller segmental and focal alterations. Copy-number analysis methods measure the abundance of genetic material and resolve these measures to genomic positions. Earlier methods were based on metaphase chromosome staining, while modern tools make use of the high-throughput capabilities and increased resolution of NGS. Likewise ASE analysis, ASCN profiling uses heterozygous variation as instruments to determine allelic skews and

copy-numbers of the two parental alleles. Heterogeneity in ploidy and purity makes the analysis challenging, but modern tools estimate these values to infer copy-number profiles suitable for between-sample comparisons. In contrast to total copy-number analysis ASCN analysis determines allelic skews and is superior in detecting LOH. Integration of ASE and ASCN analyses has the potential to reveal genomic determinants of allelic expression imbalances.

2.6.4 Quantitative trait loci

Most trait-associated SNPs, such as those identified in GWA studies, are located in the non-coding genome. Summarizing over 80 trait- and disease association studies, 88% of associated SNPs were found to lie in intronic and intergenic regions (Hindorff et al. 2009). In fact, none of the neuroblastoma risk-associated SNPs listed in table 1 is predicted to have a coding consequence[4]. Because the functional role of the non-coding genome is still not well characterized, it often remains elusive how these SNPs influence the complex traits they are associated with.

Non-coding region harbors CREs, and thus trait-associations in these regions may arise from genetic differences in regulatory elements controlling the expression of disease-relevant genes. Genetic variation explains a substantial amount of variation in gene expression between individuals (Schadt et al. 2003; Morley et al. 2004). Heritability of gene expression is the variance of gene expression explained by genetic variation. In human lymphocytes significant heritability was reported for 85% genes and median heritability was estimated to be 23% with considerable per-gene variability between 0 and 70% (Göring et al. 2007; Dixon et al. 2007). A study on Gene expression studies conducted in multiple human tissues attributed 23-36% of heritability to local genetic effects (Grundberg et al. 2012; F. A. Wright et al. 2014; Lloyd-Jones et al. 2017; Ouwens et al. 2020), suggesting that variants in gene proximal regulatory elements are an important cause for gene expression differences between individuals. Quantitative trait loci (QTLs) are individual loci associated with quantitative phenotype differences. In general, gene phenotype QTLs can be roughly classified according to the distance from their gene association: gene-associated cis-QTLs are within a fixed distance from the gene (e.g. within 1 Mb), wheras trans-QTLs exceed this distance or are located on a different chromosome. The classification helps to differentiate variants that are likely cis-acting on the gene, e.g. through a regulatory element that directly modulates gene expression of the target gene from likely trans-acting QTLs, which control

[4] according to annotations from dbSNP (build 154, released April 21 2020)

target gene expression indirectly. An example of indirect control by a tans-QTL are regulatory variants that control the expression of a transcription factor of the target gene. Cis expression (cis-e)QTLs are loci associated with a gene expression phenotype of a proximal gene. 60% of complex trait loci from GWAS were found to be linked to a cis-eQTL and associations of 18 different complex traits were found to have a median 1.7 fold-enrichment amongst cis-eQTLs (Gamazon et al. 2018), indicating cis-regulation of gene expresion as a functional mechanism underlying complex trait associations from GWA studies.

QTL analysis is a method to link genetic variation to quantitative phenotypes. It makes use of genetic and phenotypic data of many individuals to infer significant associations between genetic variants and quantitative traits, such as gene expression. In cis-QTL analysis, the set of variants associated with the quantitvaitve trait is restricted to a cis-window relative to the coordinates of the gene considered, focusing on associations of (likely) cis-acting variation on quantitative gene phenotypes (Stranger et al. 2005; Schadt et al. 2008; Battle et al. 2014). cis-eQTL analysis links genotypes to gene expression of proximal genes. In cis allele-specific expression (cis-ase)QTL analysis, these variants are associated with the ASE phenotype (Battle et al. 2014).

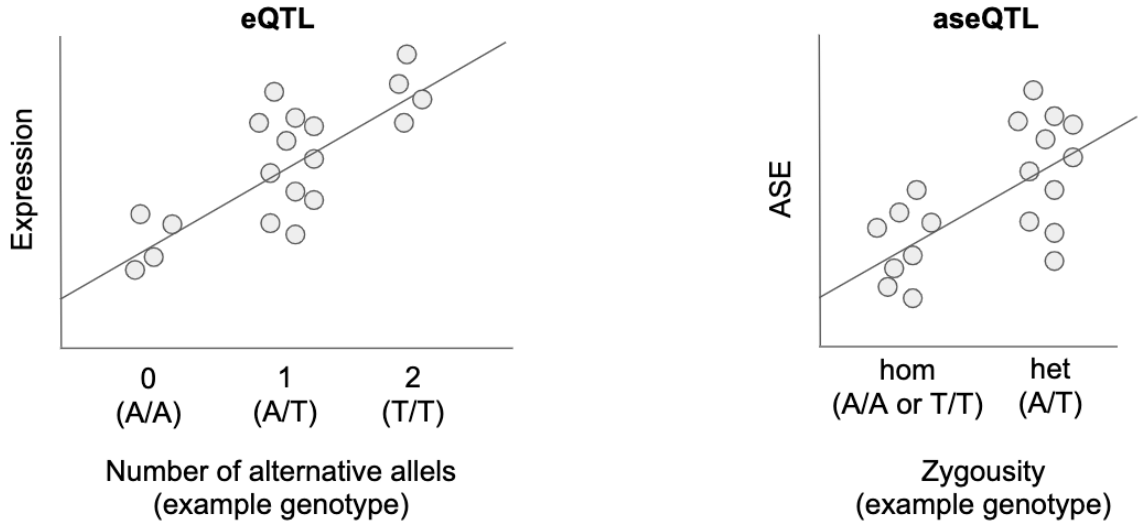


Figure 9: Regression in eQTL and aseQTL analysis. Schematic of regressions in eQTL (left) and aseQTL (right) association testing. Each dot represents a sample. In the eQTL association test the SNP genotype is encoded by the number of alternative alleles, whereas in aseQTL analysis the SNP genotype is encoded by its zygosity.

A common way to associate a single SNP genotype with a quantitative trait in QTL analysis is by linear regression (Shabalin 2012):

$$g = \alpha + \beta s + \epsilon, \text{ where } \epsilon \sim i.i.d.\ N(0, \sigma^2) \quad (4)$$

Here, *g* is the quantitative phenotype, α is the intercept, β is the slope coefficient, *s* is the genotype variable and ε the error term. The variables can be estimated by minimizing the sum of squared residuals. A test statistic is employed to determine a p-value for the association. In eQTL analysis the genotype variable can be set to the number of alternative alleles at the SNP position (0, 1 or 2). In aseQTL association the genotype is encoded by the two possible states of zygosity (homozygous, heterozygous). The ASE ratio trait quantifies the expression imbalance between alleles and this imbalance is expected to be strong for a causal heterozygous variant (e.g. a cis-regulatory variant) and weak for a homozugous variant. Therefore in the aseQTL regression model the genotype variable *s* is set to the SNP zygosity and becomes binary. Figure 9 depicts the differences between regressions in eQTL and aseQTL association tests.

Confounding factors, such as population structure, environment effects and somatic alterations can bias trait associations. In association studies systematic ancestry differences can lead to spurious associations (Kittles et al. 2002; Freedman et al. 2004). Kittles et al. found that allele frequencies of prostate-cancer associated loci differed significantly between populations; and after controlling for these differences using unlinked marker loci, the association did not reach significance. Besides unlinked marker loci (J. K. Pritchard and Rosenberg 1999), the use of principal components was suggested to model ancestry differences in association studies (Price et al. 2006). In addition to population structure, environmental and technical variability are important confounders of gene expression, which can bias QTL analysis (Plagnol et al. 2008). Because the source and structure of these confounders is often unknown and cis-QTLs are defined by their effect on the local quantitative trait, broad variance components over multiple quantitative traits are expected to represent confounding effects in cis-QTL mapping. Variance components can be inferred from expression data and resulting covariates included in association models to improve QTL mapping (Stegle et al. 2012). In addition to biases introduced by population structure, QTL analysis in cancer is affected by expression differences caused by local DNA copy-number dosage effects: Copy-number gains and losses influence gene expression traits and induce phenotypic variance that may hinder identification of QTL SNPs. However,

per-gene copy-number status can be used as model covariates, making these models better applicable to QTL analysis in cancer (Q. Li et al. 2013; PCAWG Transcriptome Core Group et al. 2020).

In the simple linear regression QTL model of equation 4, covariates can be introduced as additional terms. In the recent past, efficient association methods based on linear mixed models were developed (Christoph Lippert et al. 2011; C. Lippert et al. 2014). Here, similar to simple linear regression models, covariates can be introduced as fixed effects. However, an advantage of linear mixed models is that they can account for the relatedness between individuals by the covariance structure of random effects; and this allows for a better control of spurious associations caused by unequal relatedness between individuals (Yu et al. 2006; Z. Zhang et al. 2010).

The linkage of variant genotypes aids in the detection of QTLs, but complicates identification of causal cis-regulatory variants. In meiosis, alleles of genetic variants do not segregate to the haploid daughter cells randomly. Crossing-over and subsequent separation of homologous chromosomes result in a non-random co-segregation of alleles. Alleles of variants that are close to each other have a higher likelihood to segregate together resulting in correlation of genotypes of neighboring variants in a population, referred to as linkage disequilibrium (LD). Because of this phenomenon, non-causal variants linked to causal variants can still help to localize trait associations (N. E. Morton 2005). This even holds true if the causal variant is unknown. If, for example, the causal variant is structural and not genotyped, non-causal, genotyped SNPs in LD with the causal variant may still allow to detect an association at the locus. However, in QTL analysis, this also implies that QTLs must not be regarded as causal variants per se. Further evidence has to be collected to assign a functional role to QTLs or variants linked to these QTLs. Evidence for a functional role in cis-regulation could be an overlap of the variant with epigenetic annotations compatible with regulatory elements or functional experiments probing the variant sequence for its regulatory potential.

Complex traits were found to be associated with eQTLs. A study in human adipose and blood tissue found 50% of gene expression traits in adipose tissue to be correlated with clinical obesity-related traits (Emilsson et al. 2008). GWAS SNPs were found to be overrrepresented in cis-eQTLs of immunity-related traits in lymphoblastoid cell lines (Nica et al. 2010). An eQTL study in primary B cells and monocytes found that half of 548 complex

traits from GWAS considered were associated with one or more cis-eQTLs (Fairfax et al. 2012). The integration of eQTLs from 427 human liver samples with disease trait associations prioritized candidate susceptibility genes for CAD and LDL cholesterol levels as well as type 1 diabetes (Schadt et al. 2008). These studies show that complex trait associations are linked to gene expression traits and that combining complex traits with eQTL associations may uncover genes involved in the biological mechanisms underlying the complex traits.

eQTL studies in cancer link risk loci to cellular gene expression traits. Recently, the collection of germline variation and tumor gene expression in numerous cancer types by TCGA facilitated cancer QTL studies. The integration of GWAS and TCGA data linked cellular traits to cancer risk. At 15 breast cancer risk loci three significant cis-eQTL associations were found, priorizing IGFBP5, C5orf35 and TOX3 as candidate breast cancer risk genes (Q. Li et al. 2013). In their work Li et al. estimated that cis-acting eQTLs accounted for 1.2% of total variation in tumor gene expression. cis-eQTLs mapping in high-grade serous epithelial ovarian cancer identified loci conferring disease risk through regulation of genes HOXD9, CDC42 and CDCA8 (Lawrenson et al. 2015). A targeted eQTL mapping approach at cancer risk loci identified cis-regulation of the ABHD8 to underlie mechanisms of breast and ovarian cancer risk (Lawrenson et al. 2016). A study across five different cancer types assigned gene expression traits to 42 of 149 cancer risk loci by eQTL mapping (Q. Li et al. 2014). A recent study based on whole-genome sequencing compared cis-eQTLs across 27 tumor types with those reported by the GTEx consortium and found 12% of eQTLs were exclusively found in cancer samples, but not in GTEx post-mortem tissue samples (PCAWG Transcriptome Core Group et al. 2020). These findings suggest that cis-regulation by germline variants predisposes to cancer risk and that combining genome-wide associations with QTL analysis can prioritize cellular gene expression phenotypes underlying these risk mechanisms. To identify the correct cellular phenotypes it may be required to specifically analyze tumor samples, because some of the QTL associations discovered in cancer were not found in other somatic tissues.

By connecting complex trait associations with gene phenotypes QTL analysis uncovers genes involved in the biology of health and disease. Most complex trait associations are located in the non-coding genome, where also CREs are found. Variants in gene proximal regions are important determinants of gene expression heritability and cis-QTLs analaysis helps to identify genetic variants associated with quantitative gene phenotypes. In QTL

analysis confounders such as relatedness between individuals and somatic copy-number variation (in cancer) need to be controlled for. LD between variants helps to uncover associations but complicates the identification of the exact causal variants.

2.7 Research objectives

Somatic SVs and CN variation contribute to genetic regulation. Strong copy-number increases of DNA sequences can lead to upregulation of gene expression due to dosage effects. Several key findings in neuroblastoma indicate the importance of this form of genetic control of gene expression in the deregulation of disease-associated pathways. It is well established that DNA amplifications upregulate MYCN (Bordow et al. 1998) and also cases of ALK amplifications were described (Y. Chen et al. 2008; Schulte et al. 2011). Somatic SVs disrupt genes, lower their expression and were predicted to specifically deregulate pathways involved in neuronal development and activity (Molenaar, Koster, et al. 2012). Conversely, somatic SVs in non-coding regions can expose genes to aberrant cis-regulatory environments. Rearrangements translocate strong enhancer elements to the TERT promoter and thereby activate its expression in order to maintain telomere elongation (Peifer et al. 2015; Valentijn et al. 2015). Furthermore, chromosomal and segmental CN alterations are frequent in neuroblastoma tumors that usually harbor only low numbers of somatic mutations in protein coding genes compared to other cancers, specifically those of adulthood. Perhaps, neuroblastoma is a cancer entity that is mainly driven by CN alterations. Indeed, larger CN losses and gains on specific chromosome arms were found to impact gene expression locally (Bordow et al. 1998; Łastowska et al. 2007; Schulte et al. 2011). Somatic SNVs in non-coding regions were associated with deregulation of genes in other cancer types, most prominently in the case of TERT promoter mutations. These findings show how somatic alterations add a genetic regulatory layer to the underlying germline regulatory background. However, little is known on how strongly different classes of variation impact gene expression in neuroblastoma. For example, it has not been established if differences in the germline or copy-number effects show stronger contributions in gene expression variance in this disease. So far, mechanistic insights in genetically deregulated pathways were gained from focal alterations (such as amplifications and SV breakpoints) that could be linked to the control of individual genes. But little is known on how larger CN alterations affect pathways and disease mechanisms, such as telomerase maintenance.

To address this gap somatic and germline variation in patients and their neuroblastoma primary tumors will be identified and associated with local differences in gene expression

and disease phenotypes. More precisely, SNPs, somatic CNs, SVs and SNVs will be identified and associated with gene expression and ASE. To specifically enrich local regulatory effects (as opposed to effects in trans) I will obtain ASE and ASCN profiles in these tumors. And I will estimate the global impact on expression variability of genetic regulators and prioritize the strongest regulators for association with selected disease phenotypes. Specifically, the aim here is to investigate how cancer-associated genetic deregulation affects disease-specific survival and telomere maintenance. The results of these analyses will be presented in chapter 3.

Due to their role in MYCN oncogene amplification, ecDNAs are of particular interest in the genetic characterization of neuroblastoma tumors. Previous studies have shown that ecDNAs are associated with oncogenic amplifications of different driver genes across a wide spectrum of cancer types (Turner et al. 2017). Allele-specific analysis of ecDNA in glioblastoma showed that these molecules induce high expression from the amplified allele (S. Wu et al. 2019). Recent advances in identification of ecDNAs introduced by the Circle-seq method allows for genome-wide identification of circularized DNA and studies that applied this technique showed that circular DNA is also prevalent in healthy tissues (Henrik Devitt Møller et al. 2018; Henrik Devitt Møller 2020). These studies suggest that gene amplifications in cancer are frequently caused by ecDNA and that circularization of DNA is a common phenomenon in healthy and malignant cells. Previous work in neuroblastoma has focused on the expression of amplified double minutes, specifically the MYCN amplicon. Yet the relationship between circularized alleles, copy-number and transcription levels has not been subject to a genome-wide analysis. Circle-seq was used to study eukaryotic cells (Henrik D. Møller et al. 2015) and human somatic tissues (see above). However, so far it has not been applied to characterize circular DNA and its allelic identity in neuroblastoma tumors. Thus, the genetic regulatory role of many ecDNAs in neuroblastoma remains unexplored.

To shed light on the interplay between circular DNA, somatic copy number and gene expression, I will investigate the relation between ecDNAs, copy-number imbalances and allelic expression differences. To this end I will first establish the allelic origin of ecDNAs and determine both allelic copy-number and allele-specific expression in circularized regions of the tumor genome. To understand how circle haplotypes relate to copy-number I will quantify the abundance of circle haplotypes relative to the abundance of genomic DNA. To see if circles of certain length have characteristic copy-number patterns in the underlying genomic

DNA the length of circularized regions will be correlated to the somatic copy-number state of the circularized genomic region. To investigate how circular DNA affects gene expression dependent on the copy-number state of the circularized allele I will investigate how allele-frequencies of Circle-seq, WGS and RNA-seq are associated. Results of these analyses will be covered in chapter 4.

Several neuroblastoma susceptibility loci were identified by a series of GWAS (see Table 1). Because the identified germline variants lie in non-coding regions of the genome it is expected that they confer their effect through cis-regulation, as shown for the risk-associated intronic LMO1 enhancer SNP (D. A. Oldridge et al. 2015). cis-eQTL analysis integrates germline genotypes and gene expression quantification and has the potential to uncover cis-regulatory effects. Similarly, aseQTL mapping is based on ASE and may improve cis-eQTL mapping by reducing the influence of trans effects. As in other genotype associations, identification of functional variation in cis-QTL mapping is complicated by LD. This makes it challenging to distinguish true functional effects from those introduced by genotype correlations. Characterizing cell type-specific epigenetic properties of the DNA sequence by e.g. histone ChIP-seq or ATAC-seq (chromatin accessibility) may help to pinpoint those cis-eQTL variants that have a functional role due to their overlap with CREs, where they could alter TF binding. cis-eQTL analysis has already been applied to map SNPs to nearby expression traits in several cancers of adulthood (Q. Li et al. 2014; Lawrenson et al. 2016; PCAWG Transcriptome Core Group et al. 2020) but to my knowledge not to any childhood cancer or to neuroblastoma specifically.

To fill this gap cis-QTL mapping of two expression traits in neuroblastoma primary tumors will be conducted: WGS derived germline SNP genotypes will be associated with both total expression and ASE of proximal genes controlling for copy-number induced effects and other confounders. Furthermore candidate functional variants will be prioritized by epigenetic observations from H3K27ac ChIP-seq and ATAC-seq in neuroblastoma cell line SH-SY5Y. To investigate cis-regulatory effects, I will compare the obtained cis-QTL maps with existing GWAS summary statistic on neuroblastoma susceptibility (McDaniel et al. 2017). This analysis and the corresponding results will be presented in chapter 5.

3 Genetic effects on expression variability and disease-associated gene regulation

In this chapter genetic and cis-regulation will be related to different types of variants and their regulatory effects associated with disease traits. To determine genetic causes of deregulation of gene expression in neuroblastoma, first germline and somatic variation will be identified in 116 neuroblastoma primary tumors. I will then associate this variation with total gene expression and ASE. Using these associations I will provide estimates for the global contribution of genetic factors to gene regulation in neuroblastoma. Furthermore, I will relate ASE to underlying ASCN profiles as well as total gene expression to CN in order to identify genes and pathways that are subject to dosage-dependent deregulation by these larger variants. Genetic regulatory effects will be linked to survival and telomere maintenance mechanisms. And correlation of ASE and total gene expression of differentially expressed genes will be used to identify genes of consistent disease-associated deregulation in cis.

Contributions to this chapter

Alignments of normal WGS, tumor WGS, tumor RNA-seq and somatic single nucleotide variant calls were created by the Core Unit Bioinformatics (CUBI) of the Berlin Institute of Health (Berlin, Germany) under supervision of Dr. Dieter Beule. Dr. Jörn Tödling (AG Schulte, Charité Universitätsmedizin Berlin, Germany) provided somatic structural variant calls of neuroblastoma primary tumors. Remo Monti (AG Ohler, MDC, Berlin, Germany) integrated the fastlmm and PEER methods into the data processing pipeline that were used to map cis-eQTL and cis-aseQTLs. Christiane Weber performed an initial ASCN-survival association test on a subset of samples (Peifer et al. 2015) that provided valuable input for the analysis and interpretation of results described in sections 3.1.8 and 3.2.9.

3.1 Methods

Our analysis integrates WGS of matched normal and tumor tissue and RNA-seq of tumor tissue with a variety of external data sources, including SNPs and genotypes from the broader population. In total more than 30 terabyte of NGS alignments from donors of neuroblastoma tumors were processed. For this purpose a distributed compute environment was used to parallelize execution of many processing steps. We implemented a pipeline that

produced the data files for the downstream analyses presented in this and subsequent chapters. Among other tasks the pipeline performs genotyping, phasing, ASCN calling, quantification of gene-level ASE and total expression, as well as cis-eQTL and cis-aseQTL analysis (see Figure 10a). Many of the processing steps are interdependent. For example, allele-specific analysis of DNA and RNA requires identification of hetSNPs from the genotyping step to determine allelic read counts of tumor and normal WGS (for ASCN) as well as RNA-seq (for ASE) (see Figure 10b). We made use of the workflow management snakemake (Köster and Rahmann 2012) (version 5.9.13), a tool that resolves such dependencies and executes defined processing steps in the correct order. We will not provide an exhaustive description of our pipeline here, but will include the description of critical processing steps together with techniques used in the downstream analysis in the corresponding method sections of each chapter. The pipeline program code is provided as supplementary material on the data storage attached to this thesis.

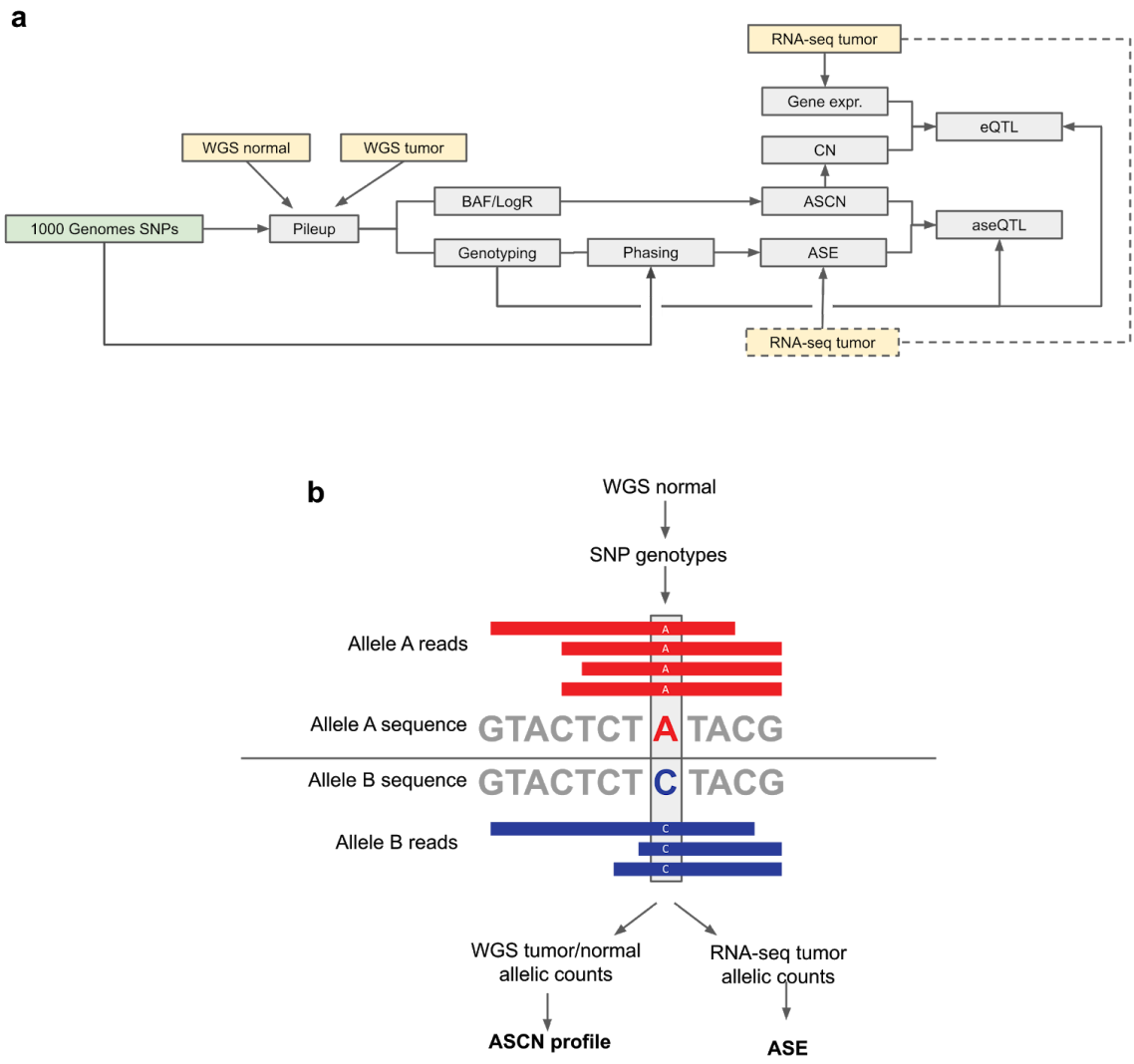


Figure 10: Data processing pipeline and allele-specific readouts. **a**, Schematic overview of the data processing pipeline with dependencies between selected processing steps. Yellow: input samples from neuroblastoma patients. Green: SNPs and genotypes in the broader population. **b**, Allelic counts at heterozygous SNPs in WGS and RNA-seq alignments are used to infer allele-specific copy-number profiles and allele-specific expression respectively. Depicted in (b) are sequences of two alleles differing at a single nucleotide, the heterozygous SNP with genotype A/C. Reads from alignments of WGS and RNA-seq spanning this heterozygous SNPs are assigned to one of the two alleles (indicated in red and blue) by their agreement to one of the two alleles at the SNP position. WGS: Whole-genome sequencing, BAF: B-allele frequency, CN: copy-number, ASCN: allele-specific copy-number, ASE: allele-specific expression. eQTL: expression quantitative.

3.1.1 Sample preparation and sequencing

Samples were collected in the NB2004 trial between 2004 and 2016 in a cooperative multi-centric study in the university hospitals of Cologne and Berlin. DNA and RNA samples of 38 donors were obtained from primary tumors of at least 60% tumor cell content as evaluated by a pathologist. MYCN copy-number was determined by FISH in clinical routine diagnostic. Sample preparation and sequencing was performed as described earlier (Koche et al. 2020). Briefly, WGS of tumor-normal pairs was performed on the HiSeq X-Ten platform (Illumina, San Diego, USA), yielding paired-end reads of 2×150 bp length. Ribo-depleted RNA was sequenced on the HiSeq4000 platform (Illumina, San Diego, USA) yielding reads of 2×150 bp length. Additional sequencing data was obtained from the European Genome-phenome Archive[5] under accession number EGAS00001001308 for a non-overlapping set of donors from a previous study on somatic structural rearrangements in neuroblastoma (Peifer et al. 2015). After quality control 52 donors of this study were included, yielding a total of 116 donors with matched tumor RNA-seq, tumor WGS and blood-derived normal WGS.

All reads were aligned to the GRCh37 (hg19) reference. WGS reads were aligned with BWA-MEM 0.7.15 (H. Li and Durbin 2009). RNA-seq reads were aligned with STAR 2.5.3a (Dobin et al. 2013). Samblaster 0.1.24 (Faust and Hall 2014) was used to mark duplicates in alignment files. Quality control was performed using FastQC[6]. Supplementary table 1 lists samples and donors, from which sequencing data was obtained and used in the analyses.

3.1.2 Telomere length analysis

Telomeres length was estimated from WGS of normal and tumor samples by Telseq 0.0.2 (Ding, Mangino, et al. 2014) with parameter -u (ignore read groups) and otherwise default settings. Briefly, the method estimates telomere length by counting WGS reads containing the telomere repeat sequence $(TTAGGG)^k$, where k denotes the number of repeats of the 6-mer. Telseq uses default repeat length k = 7 and normalizes the resulting read count by GC content and a genome size factor. The authors calibrated the default parameters using telomere length measurements determined by southern blot analysis of terminal restriction fragments. We summarized telomere lengths per sample by the log telomere length ratio

[5] https://www.ebi.ac.uk/ega/, accessed 18 Mar 2021
[6] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed 18 Mar 2021

$\log(L_T/L_N)$, where L_T and L_N are the Telseq estimates for telomere length in tumor and normal WGS sample respectively.

3.1.3 Total gene expression analysis

Aligned tumor RNA-seq reads were counted using HTseq/htseq-count 0.9.1 (Anders, Pyl, and Huber 2015) on exons of protein coding genes according to Ensembl[7] release 75 human gene annotations for the GRCh37 reference, summarizing counts on gene-level. DESeq2 1.26.0 (Love, Huber, and Anders 2014) was used to perform differential expression analysis between donors marked as deceased from disease according to the clinical annotation file and other donors. Six donors were excluded from the analysis, because of missing clinical annotations. p-values and log-fold changes of differential expression were obtained controlling for sample covariates cohort, tumor purity, age and sex. Log-fold changes were shrunken using the apeglm method (Zhu, Ibrahim, and Love 2019). Pathway enrichment analysis was conducted using the fgsea R package (Korotkevich, Sukhov, and Sergushichev 2019) version 1.12.0 with inbuilt Reactome pathway definitions. For other analyses than the differential gene expression described above we normalized gene expression for the purpose of between-sample comparisons in a given gene. To mitigate the effect of sequencing depths and batch effect introduced by different RNA library preparation- and sequencing methods between the two cohorts we normalized htseqs by the following strategy: We first calculated library-size normalized DESeq2 variance stabilized counts from htseq counts. Then, we modeled the variance stabilized counts by cohort membership using simple linear regression for each gene and determined the residual for each gene and sample. If not indicated otherwise, this residual was used as the measure for total gene expression in our analyses.

3.1.4 Genotyping and phasing

Variant call files with 84,801,880 germline variants reported by the 1000 Genomes Project (phase 3) (1000 Genomes Project Consortium et al. 2015) were downloaded and filtered for biallelic SNPs. SNPs from chromosome 1-22 were filtered for minor allele frequency (MAF) of 1% or higher in the 1000 Genomes cohort. A mappability signal track (wgEncodeCrgMapabilityAlign50mer) for 50-mers in the human reference hg19 was downloaded from the UCSC genome browser (Haeussler et al. 2019) and intersected with the SNP positions. Only SNPs with a mappability score of 1 (unique 50mer) were kept. The

[7] http://www.ensembl.org/, accessed 18 Mar 2021

resulting set of 9,866,569 variant sites was defined as the *SNP panel* for further downstream analysis.

Pileups are data structures that make nucleotide base observations in sequence alignments easily accessible. For a given alignment pileups can be created for a subset of covered positions. A pileup reports the counts and identity of aligned nucleotides, mismatches and alignment gaps for each position considered. This information can be used to analyze allelic frequencies and base quality scores at predefined genomic positions, such as those of previously known SNPs. In DNA samples of healthy tissue (e.g. blood) the information of the pileup can also be used to assign genotypes to SNP positions. We generated pileups at positions of the SNP panel from whole genome sequencing (WGS) alignments of blood-derived control samples by Bcftools 1.8 mpileup[8], excluding unmapped reads, or reads that were marked as optical duplicates or “not primary alignment”. The resulting pileups were then used as input to the Bcftools 1.8 multiallelic-caller to call genotypes at the positions of the SNP panel. Briefly, the caller determines the number of observations per allele and employs a statistical model incorporating these frequencies and read quality scores to determine genotypes (homozygous reference, homozygous alternative, heterozygous) and a genotype quality score. We only kept resulting genotypes with an allelic depth of 10 or more reads and a genotype quality of 20 or higher. The resulting individual variant files were merged and genotypes were phased by Eagle 2.4 (Loh et al. 2016) using the phased 1000 Genomes genotypes as reference. In this step each of the two alleles (reference and alternative) is assigned to an A or B allele, which represent the two parental alleles. The method is based on dependencies in allele frequencies of neighboring SNPs and makes use of pre-existing information of haplotype assignments in the reference panel (here the phased 1000 Genomes cohort). The resulting variant file, comprising phased genotypes of all individuals was defined as the *genotype panel* for further downstream analysis.

3.1.5 Allele-specific expression analysis

Allele-specific RNA read counts were determined by GATK (McKenna et al. 2010) (version 3.5.0) ASEReadCounter from RNA-seq alignments at heterozygous SNPs following established protocols (S. E. Castel et al. 2015). Sites (i.e. SNP in a sample) with less than 8 total or less than 2 allelic reads were removed. Additionally, only sites that qualified as bi-allelic according to a statistical test were retained: A binomial test on the minimum allele

[8] http://www.htslib.org/doc/bcftools.html

count = min(alt, ref), number of trials (alt + ref) and hypothesized probability of success sum(non_ref_alt)/sum(raw_depth) was applied, where ref and alt are the reference and alternative allele counts, and non_ref_alt and raw_depth the non-reference/non-alternative allele count and raw read depth per site respectively. Sites for which the null hypothesis was rejected (FDR 0.05, Benjamini-Hochberg) were classified as bi-allelic. The reference allele bias was estimated by averaging over the reference allele fraction ref / (ref + alt) of all ASE sites from balanced copy-number regions per sample. We used statistical phasing information (Section 3.1.4) to summarize allelic counts at exonic hetSNPs of the same haplotype per gene. Only genes with a total of 10 or more counts from both haplotypes were retained. The ASE ratio for a given gene was calculated as max(A, B) / (A + B), where A and B are haplotype counts of the arbitrary A and B allele respectively. Expression imbalances per gene and sample were assessed by a two-sided binomial test using A as the number of successes, (A + B) as the number of trials and 0.5 as the hypothesized probability of success. The p-value was adjusted for multiple testing using the Benjamini-Hochberg procedure. Allelic-expression imbalance (AEI) status was assigned to observations (gene-sample pairs) for which an expression imbalance was detected at FDR 0.05.

3.1.6 Allele-specific copy-number analysis

Pileups of primary tumor WGS were generated by Bcftools 1.8[9] mpileup at SNP positions of the genotype panel established in section 3.1.4. Unmapped reads, or reads that were marked as optical duplicates or as “not primary alignment” were not considered in the pileup. For each of the SNPs the allelic depths were calculated from the pileups on normal (see ref:Genotyping_and_phasing) and tumor alignments respectively. For SNPs with total depth of 10 or more reads in both tumor and normal alignments we determined the B-allele frequency (BAF) and the coverage log ratio (LogR). For a given pileup position the BAF is defined as the ratio between alternative allele nucleotide count and the number of total considered counts $a_i/(r_i+a_i)$, where a_i and r_i are the allelic depths of alternative and reference allele respectively. The LogR at SNP position *i* was defined as $\log2((d_{ti}/d_{ni})/(\partial_t/\partial_n))$, where d_{ti} is the total depth at SNP position *i* in the tumor sample, d_{ni} is the total depth at SNP position *i* in normal sample and ∂_t and ∂_n are mean depths at SNPs of tumor and normal sample respectively.

The BAF of a heterozygous SNP position is informative for the proportion of aligned reads originating from the paternal and maternal allele. At a homozygous SNP the BAF is

[9] http://www.htslib.org/doc/bcftools.html, accessed 18 Mar 2021

expected to be close or equal to 1, if the sample's SNP genotype is homozygous alternative or close or equal to 0 if the genotype is homozygous reference. The BAF is calculated separately for alignments of normal and tumor, resulting in a *normal BAF* and a *tumor BAF* per SNP and sample. The LogR is a measure of total coverage difference between normal and tumor samples and is informative at any position, including homozygous and heterozygous SNPs. It is calculated for a pair of alignments (tumor and normal), resulting in a LogR value per SNP and sample.

Allele-specific copy-number profiles were generated from tumor and normal BAFs and LogR values for each sample using ASCAT 2.6 (Van Loo et al. 2010) with a custom segmentation procedure. In ASCAT's segmentation step the BAF and LogR values are converted into intervals of similar values. ASCAT's original implementation of this segmentation considers both LogR and BAFs to obtain start and end points for segments. We found noisy coverage log ratios to introduce over-segmentation in some samples and therefore replaced the segmentation procedure with a custom implementation that only considers BAFs to determine start and end points of segments, but still estimates the segment's coverage using the log coverage ratios. ASCAT's output comprises copy-number segments with integer copy-numbers of major and minor alleles as well as estimates for tumor purity and ploidy. All CN segments were inspected manually for quality. For samples with estimated tumor purity less than 60% CN calling was rerun with adjusted purity and ploidy values that were manually selected after inspection of the goodness-of-fit plots and in agreement with pathology estimates of tumor purity (Supplementary table 2 lists manually selected purity and ploidy values for the affected samples).

Tumor purity is defined as the fraction of tumor cells in the sample or biopsy, for which the DNA was extracted. E.g. infiltration of immune cells, vascularization or capturing non-cancerous neighboring tissue in a biopsy decreases the purity of the tumor sample. Tumor ploidy describes the DNA content of tumor cells by their average haplotype count. Most healthy human cells contain two sets of chromsomes 1-22 and a pair of sex chromsomes and thefore have a ploidy of 2. In tumor cells this value can deviate due to chromosomal aberrations. Allelic gains may increase the overall DNA content and can result in ploidy values above 2. In contrast, if losses are dominating, the ploidy may be lower than 2. Because chromosomal aberrations often only affect a subset of the genome (such as individual chromosome arms or smaller regions) tumor ploidy may represent a fraction or floating point value and not necessarily an integer.

We characterized copy-number segments based on their allele-specific copy-numbers, coverage log ratio and size. We also assigned these characteristics to genes by their overlap with copy-number segments and SNP positions for local gene-level logR measurements. Copy-number states (CN state) were assigned to each segment based on logR measurements and ASCAT's allele counts and ploidy estimates. CN state *gain* was assigned to segments for which $CN_{major} + CN_{minor} > round(ploidy)$, where CN_{major} and CN_{minor} are allele counts of major and minor allele respectively and round(ploidy) the ploidy estimate determined by ASCAT rounded to an integer value. CN state *loss* was defined as $CN_{major} + CN_{minor} < round(ploidy)$. CN state *neutral* was defined as CN_{major} + CN_minor = round(ploidy). CN state *focal amplification* was assigned to segments smaller than 10 Mb with $CN_{major} \geq 5$ and $logR_{seg} - logR_{contig} > 0.7$, where $logR_{seg}$ is the mean logR of the segment and $logR_{contig}$ the mean logR of the segment's chromosome (contig). Similarly we assigned a copy-number balance state (CN balance state) to each segment. For this purpose the copy-number ratio was determined as $CN_{ratio} = CN_{major} / (CN_{major} + CN_{minor})$. Then the CN balance state *balance* was assigned if $CN_{minor} > 0$ and $CN_{major} = CN_{minor}$. CN balance state *weak imbalance* was defined as $CN_{major} > CN_{minor}$ and $CN_{ratio} \leq ⅔$, and state *strong imbalance* was defined as $CN_{major} > CN_{minor}$ and $CN_{ratio} > ⅔$. CN balance state *LOH* was assigned if $CN_{minor} = 0$. CN balance state *amplification* was defined in the same way as for the CN state above. In addition to the state of CN segments we also assigned CN states to genes. For this purpose the overlaps between gene coordinates (Ensembl version 75) and CN segments were determined. The CN state of the segment with largest overlap to the gene was assigned as the CN state to the gene. The gene's amplification status was inferred from its CN state and gene-specific logR measurements. Genes of CN state *focal amplification* or those with $logR_{gene} > 2.5$ were defined as amplified, where $logR_{gene}$ is the mean logR across all SNPs falling within the gene's coordinates.

3.1.7 Somatic single nucleotide and structural variation calling

Somatic SNVs were called by Mutect2 version 2.2 from the GATK software package (McKenna et al. 2010) using command line parameters listed in Supplementary table 3. SNV calls were filtered using a panel of normals and command line parameters listed in Supplementary table 4. Effects of SNVs were predicted using the Ensembl variant variation effect predictor version 101 (A. D. Yates et al. 2020) in offline mode with distance 100,000 bp. SNVs in categories missense, splice, stop, synonymous, 5' UTR and 3' UTR were summarized to gene level somatic mutation burden. Somatic SNVs annotated as promoter

variants by the Ensembl variant effect predictor were considered separately. Splice, nonsense and missense variants for each gene were summarized based on the assigned consequence.

SV were called using the software novobreak version 1.1.3 (Chong et al. 2017) in pairs of matched tumor and normal WGS alignments. Briefly, novobreak detects SVs by analyzing k-mers that are unique to reads in the tumor sample. K-mers of tumor reads are collected and those occurring in normal samples or in the reference sequence are removed. The remaining k-mers are clustered and local assemblies of identified kmer-containing reads are generated. Consensus sequences of the assemblies are then aligned to the reference genome to infer breakpoint positions. We only kept SV calls with QUAL ≥ 30, at least 5 high quality reads in support of each breakpoint in the tumor sample, 0 reads supporting each breakpoint in the normal sample, 5 or more discordant reads per breakpoint in the tumor sample and 3 or less discordant reads per breakpoint in the normal sample. The functional effects of SVs at the TERT locus have been established previously (Peifer et al. 2015; Valentijn et al. 2015) and for the detection of TERT SVs we relaxed the threshold on high quality reads in support of each breakpoint, requiring at least 2 of those reads to keep the SV call. Other thresholds were applied as described above. TERT rearrangement status was assigned to a sample positive for at least one somatic SV 100,000 kb upstream or downstream from TERT gene start and end coordinates (Ensembl/GRCh37) or annotated as TERT rearranged in Peifer et al. 2015.

We used a targeted approach to identify ATRX exon deletions. To this end we determined read coverage at ATRX gene coordinates in 50 bp bins, normalized the read counts by the number of overall mapped reads and defined a tumor coverage ratio by $s_i = \log2(n_{iT}/n_{iN})$, where n_{iT} and n_{iT} is normalized read count in tumor and normal for bin i respectively. For each matched tumor/normal pair we then fit a two-component gaussian mixture model to the signal and determine the mean and relative proportions of two hypothetical clusters, corresponding to read coverages of deleted and intact regions of the gene. Samples that harbored a signal mean difference of at least 1.5 units between the two clusters and in which the smaller cluster showed a proportion of 10% or more of the larger cluster were regarded as ATRX deleted. Tumors that showed either ATRX deletions as determined by the method our targeted approach, were found positive for a somatic SV breakpoint inside ATRX gene boundaries or carried a somatic missense, nonsense or splice SNV were considered to have an altered ATRX gene.

3.1.8 Copy-number association testing

To associate allelic copy-number differences with phenotypes we summarized copy-number ratio and logR values in genomic regions. We calculated the copy-number ratio as CN_{ratio} = CN_{major} / (CN_{major} + CN_{minor}), where CN_{major} and CN_{minor}, are major and minor allele counts as determined by allele-specific copy-number analysis respectively (Section 3.1.6). CN_{ratio} and logR values were summarized both on the level of chromosome arms as well as in 5 Mb bins along the genome. The average value per region was defined as the mean value of CN segments overlapping the genomic region weighted by the length of overlap. 5 Mb bins overlapping focal amplifications were assigned the value of the amplified CN segment directly, dropping values of other segments overlapping the same bin. We used this strategy in order to maintain copy-number signals of these small (focal) alterations independent from the choice of bin size, the size of amplified segments and the relative positioning of bins and amplified segments to each other. Note that this strategy was only applied to 5 Mb regions but not to chromosome arm-level regions.

We then associated the summarized copy-number ratio per region with patient survival. For each region we tested for the association of copy-number ratio to survival using a generalized linear regression on the binary response “deceased” vs. “not deceased”, where the “deceased” was set if the clinical status of a sample corresponded to “deceased from disease”. The test was set up to control for covariates MYCN amplification, age, tumor stage 4, sex, tumor purity and tumor ploidy. The association p-value was determined by an analysis of variance (ANOVA) using a Chi-Squared test. The test was carried out between a generalized linear model (GLM) of the covariates above and a second model that included the copy-number ratio in addition to these covariates. Nominal p-values determined for each region were corrected by the Bonferroni method and regions below 0.05 FWER were considered significant.

We used a Cox proportional hazard model (D. R. Cox 1972) to predict overall survival from the copy-number ratio of the chromosomal region identified in the regression analysis described above. In contrast to the binary outcome (deceased, not deceased), here survival times are taken into account. Subsequent significant bins in the discovery model were merged and the average copy-number ratio was determined for the merged bins by the weighted average method as described above. Survival times were predicted by the covariates copy-number ratio, MYCN amplification status, age, tumor stage 4, sex, tumor

purity and tumor ploidy. A survival function was estimated by the Kaplan-Meier method. Here discretized states “balance” and “imbalance” were used to split samples into two groups and to plot the corresponding survival curves.

We associated the copy-number logR with telomere length. For this purpose we set up a binary criterion on telomere length, dividing the samples into two groups: Samples with log(TLR) > 0.5 were assigned to the “long telomeres” group and samples with log(TLR) ≤ 0.5 to the “short telomeres” group. We then used this binary outcome as the response of GLMs using the same covariates, test- and p-value correction strategy as described for association testing between copy-number ratio and survival described above.

3.1.9 Variance component analysis

We modeled both ASE and residual total expression by local genetic effects based on detected germline and somatic variation at the respective gene locus and additional covariates using linear regression. ASE was modeled by the heterozygosity status of the SNP with greatest effect size from eQTL and aseQTL mapping (see methods 4.1.1), the copy-number ratio and binary variables indicating the presence of a structural variation breakpoint overlapping with gene coordinates including +/- 100kb flanking regions, somatic SNVs in the promoter, and at gene coordinates (including UTRs and introns) as determined by Ensembl variant effect predictor (VEP) (version 101). Similarly, residual gene expression was modeled by the genotype (encoded as number of alternative alleles) of the SNP with greatest effect size from eQTL and aseQTL mapping, copy-number logR, somatic structural variation and somatic SNVs in promoter and gene (as above). Tumor purity and MNA status were used as additional covariates in models of both expression phenotypes. In the ASE model, the log sum of coverage at the ASE SNPs was used as an additional covariate. A linear model with up to 116 observations was fitted for each gene separately. Only genes with 20 or more complete observations (for effects/covariates and expression phenotype) were considered. The explained variance per effect was determined by its relative contribution to the total sum of squares as determined by ANOVA on the fitted model. Significant variance components were determined by ANOVA’s F-statistic and the resulting p-value was adjusted for multiple testing by the Bonferroni method for each effect. Significant effects per gene were defined as effects at FDR < 5%.

3.1.10 Correlation analysis of allele-specific and total expression

To determine genes underlying strong cis-regulatory control by activation or attenuation of gene expression from one of the two alleles, we performed a correlation analysis between ASE and gene expression. ASE ratios (Section 3.1.5) were filtered, so that only ratios from 10 or more RNA-seq read counts remained. Variance stabilized read counts of gene expression were matched with ASE ratios by sample and gene. We then grouped observations by gene and only considered genes with at least 10 observations yielding 11358 genes with sufficient number of observations. Both ASE ratio and gene expression read counts were separately corrected by batch and tumor purity by fitting linear models per gene and obtaining residuals of expression and ASE ratio, that were used in the subsequent analysis. Finally the r^2 was obtained by linear regression between residuals of ASE ratio and expression counts. We defined a set candidate allelic regulated genes (AR genes) as those genes with $r^2 > 0.3$. We estimated the contribution of copy-number to the ASE ratio by linear regression of both copy-number ratio and tumor DNA ratio against ASE ratio residuals per gene. Analogous to the ASE ratio (Section 3.1.5), the tumor DNA ratio was defined as max(A,B)/(A+B), where A and B are phased and aggregated read counts of the tumor DNA alignment of expressed heterozygous SNPs gene for the two alleles respectively. To identify a subset of AR genes with clinical relevance we matched ASE-expression r^2 values with adjusted P values from differential expression analysis described in section 3.1.3. A subset of differentially expressed genes was defined by intersecting AR genes with genes significantly different expressed between deceased and not-deceased patients (FDR < 0.05, Benjamini-Hochberg).

3.2 Results

3.2.1 Germline and somatic variation in 116 neuroblastoma tumors

We characterized variation in 116 donors and their neuroblastoma primary tumors. To that extent WGS of normal tissue was used to infer germline SNPs and both WGS of normal and tumor reads and alignments were used to infer somatic alterations, including SNVs, allele-specific copy-number, and somatic structural variation.

Discovery of SNPs was restricted to those with MAF > 1% as reported by the 1000 Genomes project phase 3. We detected 9,866,569 SNPs, for which we found the alternative

allele at least once in our cohort. 83% of inferred genotypes were homozygous and 17% heterozygous. 94% of homozygous SNPs were called for the the major allele as defined by the allele with higher occurrence frequency in the cohort. We found on average 1,041,783 (893,775 –1,164,753) homozygous SNPs of the alternative allele and 1,665,192 (1,442,754 – 2,182,311) heterozygous SNPs per sample. Intersection of SNP locations with those of genomic features from the classes promoter upstream (1 to 5 Kb), promoter, 5' UTR, exon, intron and 3' UTR identified the fraction of SNP falling within the respective genomic regions. We found the vast majority (96%) of identified SNPs to be located in non-coding regions of the genome with introns and intergenic regions being the most prevalent features. 590,891 SNPs (3%) were located in exons and 269,092 (1%) in promoter regions. Figure 11 shows a summary of SNP statistics, including the distribution of minor allele frequency and heterozygosity rate per SNP as well as distribution of features overlapping SNP positions.

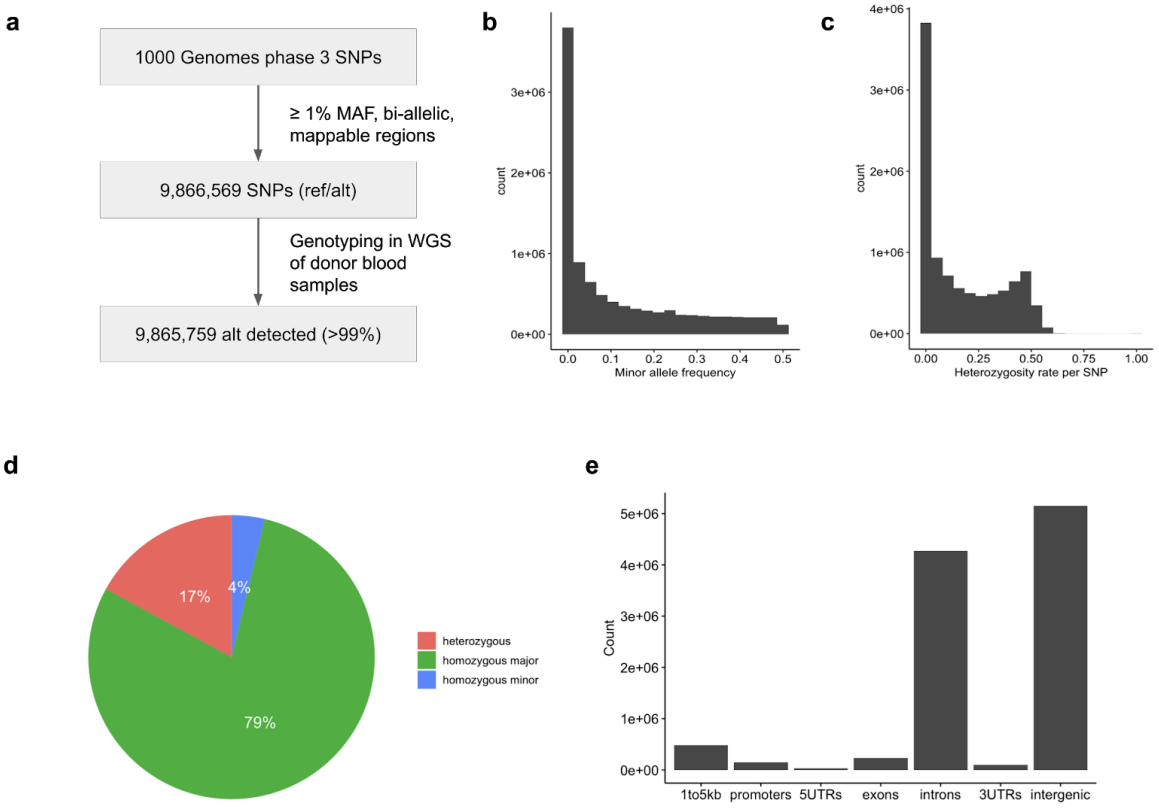


Figure 11: SNP genotypes in the neuroblastoma cohort. **a**, Genotypes were determined at bi-allelic SNP positions as reported by the 1000 Genomes phase 3 after applying filters on minor allele frequency (MAF) and mappability. **b**, Frequency of MAFs per SNP. **c**, Frequency of fraction of heterozygous samples per SNP. **d**, Distribution of homozygous and heterozygous genotypes. **e**, Number of SNPs by annotation features.

Genome wide somatic ASCN profiles were derived from BAF and logR values at SNP positions using an adapted version of ASCAT (Van Loo et al. 2010). These CN profiles consist of a series of copy-number segments and respective integer CN for the major and minor allele. Here, the major allele is defined as the allele with higher CN. The choice of major allele is arbitrary for balanced copy-number regions. We overlapped SNP-based logR measurements with these segments to infer the average logR per CN segment and together with major and minor allele count used these values to assign CN states to segments (Section 3.1.6). We found that the majority (51%) of CN segments across the tumor samples had one copy of paternal and maternal chromosomes each (neutral copy-number state), 27% were classified as gains, 21% as losses and less than 1% focal amplifications. The mean segment size across all samples was 56 Mb. However, the proportion of CN states and the distribution of CN segment sizes varied considerably between samples (Figure 12a). Samples with higher number of CN segments showed smaller segment sizes, which is expected as the sum of segment sizes per sample cannot exceed the genomic length. We determined the genome-wide frequency of CN alterations in 5 Mb genomic bins, assigning CN states to bins by those of overlapping segments. We find pronounced preferences for losses and gains in distinct genomic regions and most preferences consistent along entire chromosome arms (Figure 12b). Chromosome 17q and chromosome 7 were most frequently affected by gains, whereas chromosomes 11q, 3p and 1p harbored the most frequent losses. Highly recurrent amplifications were exclusively detected on chromosome 2p at the MYCN locus.

We investigated gains and losses in terms of logR of chromosome arms per sample. Chromosomal gains increase the relative coverage of tumor DNA compared to normal DNA alignments, whereas losses decrease the relative coverage, which is reflected by higher and lower logR values respectively. We determined the average logR across all segments overlapping a chromosome arm, and then clustered and compared the resulting patterns across samples. We find substantial between-sample heterogeneity reflected by diverse patterns of chromosome-arm level logR across the investigated tumors. High logR of chromosome arm 17q is found among all three risk groups. However, certain gains and losses are more prevalent in high and low risk groups respectively: Low 11q logR is more prevalent among high risk samples, whereas most low risk samples show high logR values of chromosome 7. Samples harboring MYCN-amplifications cluster into two different groups: The first group shows smaller absolute logR values compared to other high risk samples, indicating the absence of copy-number alterations on most chromosomes in these samples.

The second group of MYCN-amplified samples comprise samples with heterogeneous logR patterns involving losses of 11q and chromosome 9, losses of 17p and chromosome 10. Both groups show prevalent losses of 1p, an alteration known to be linked to MYCN amplification in neuroblastoma (Fong et al. 1989) and share 17q gains with other high- and low-risk tumors. Figure 13 shows the chromosome arm logR for each sample grouped by risk stratification and annotated with additional clinical and genomic features.

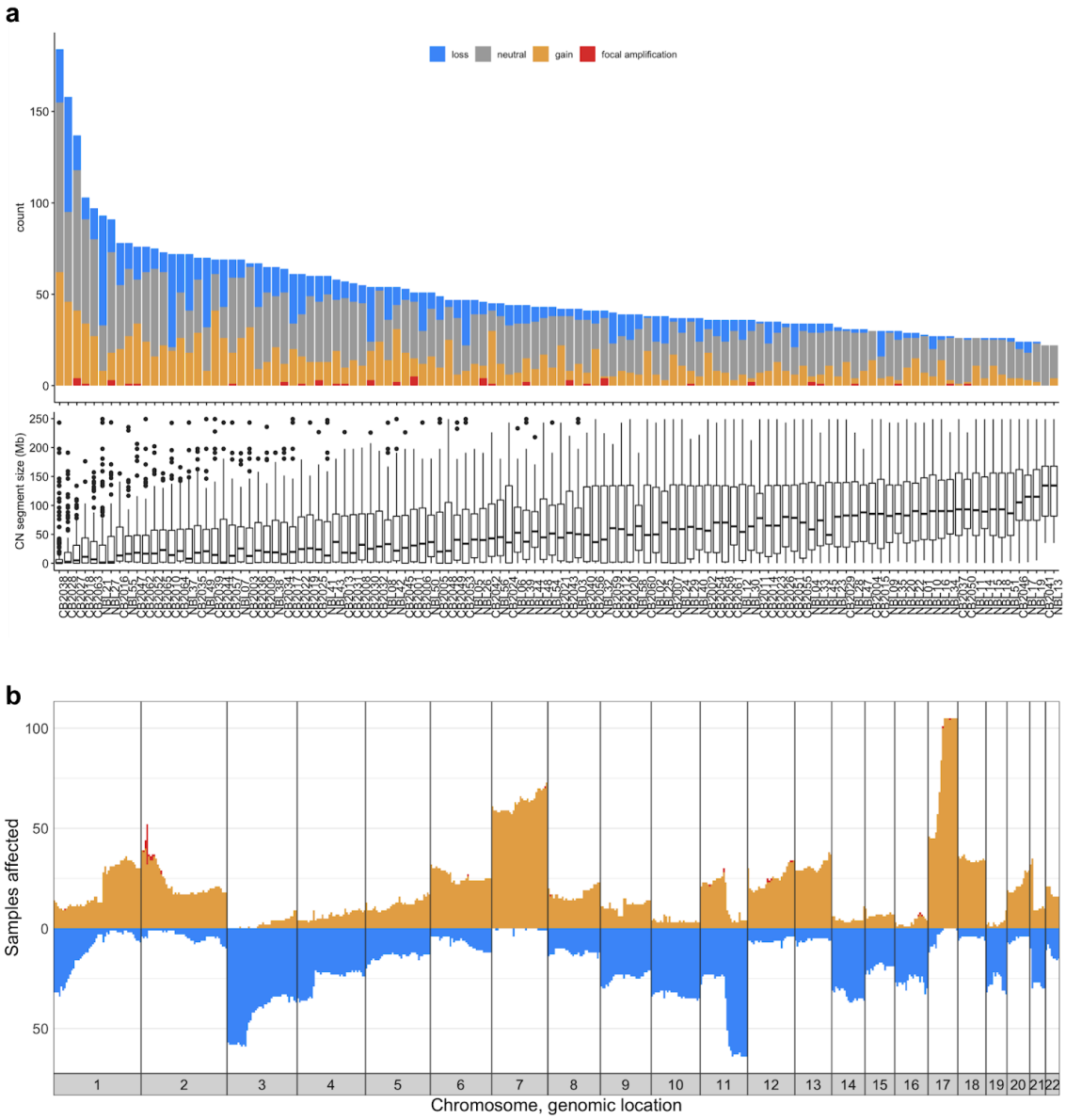


Figure 12: Copy-number segmentation and states across 116 neuroblastoma tumors. **a,** Number of copy-number segments by copy-number state (top) and the distribution of copy-number segment sizes (bottom) for each sample. **b**, Number of samples affected by copy-number status changes summarized in 5 Mb genomic bins. Yellow/red: Number of samples affected by gains and amplifications, Blue: Number of samples affected by losses (reverse scale).

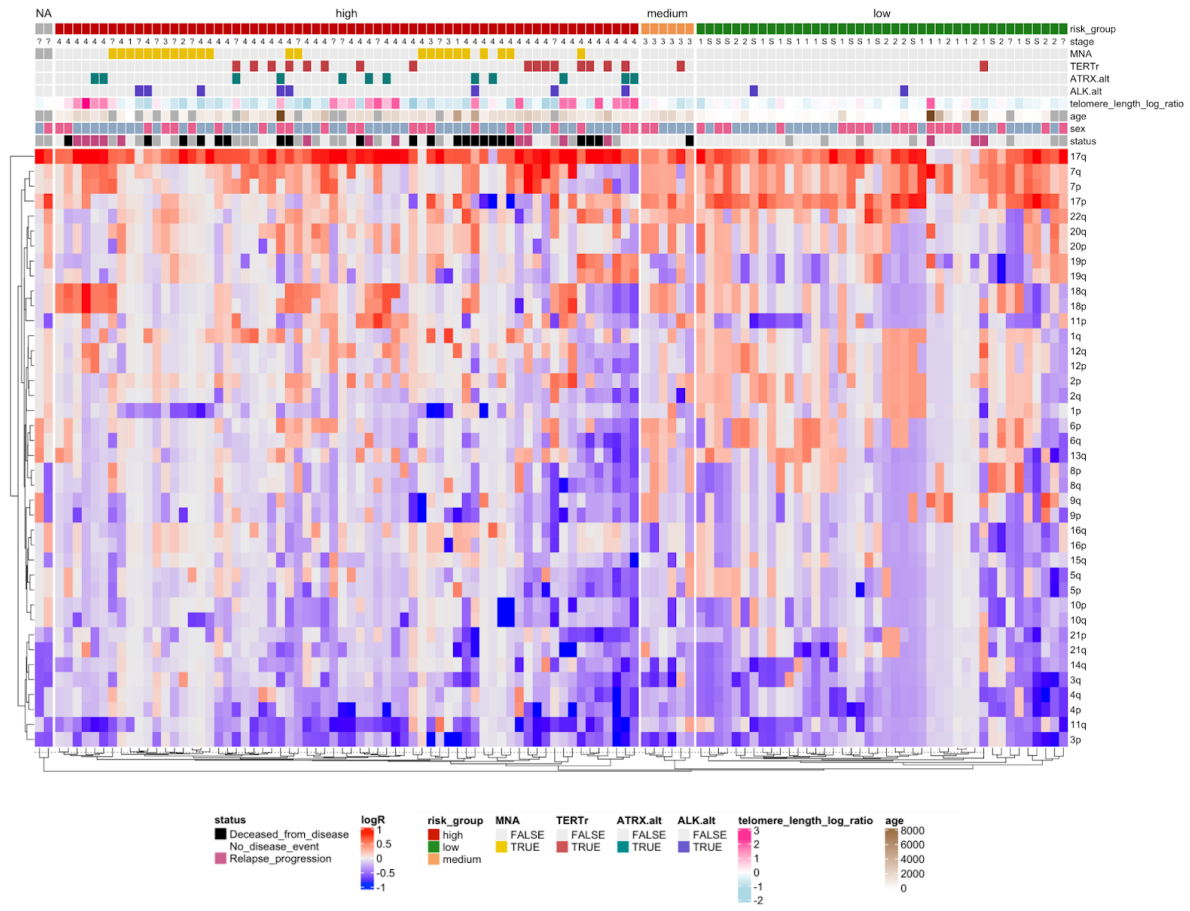


Figure 13: Log ratio of tumor and normal coverage per chromosome arm. Each column represents a tumor sample. MNA: MYCN amplification, TERTr: TERT re-arrangement. ATRX.alt: ATRX alteration, ALK.alt: ALK alteration, logR: log ratio.

We defined focal amplifications as small segments with very high tumor DNA coverage (Section 3.1.6). To investigate genes that are affected by these strong copy-number increases, gene coordinates were overlapped with coordinates of focally amplified segments. Overlapping genes were then marked as amplified. Additionally we determined a logR value per gene and marked those genes as amplified that showed extreme logR within their coordinates (Section 3.1.6). We find 32 samples to harbor gene amplifications and a total of 357 genes to be amplified across all samples. Figure 14 shows the number of amplified genes per samples, their genomic location and predicted interactions between recurrent ampliciations. MYCN amplifications were confirmed in 20 of 24 samples that were annotated as carriers of the amplification in the clinical data. Notably we did not find evidence for MYCN amplifications in tumors of samples CB2045, CB2031, CB2045 and CB2047, despite their amplification status in the clinical annotation. Interestingly, among these samples CB2045 shows a focal loss at the MYCN locus instead within a broader region of LOH (Figure 36). We speculated that amplifications may target established cancer

driver genes and determined which of the affected genes are part of the set of cancer census genes from the COSMIC database[10]. We find the following cancer consensus genes among the amplified genes: ALK, BCL7A, BRCA1, CCND1, CDH1, CDK4, CLIP1, ETV4, LRIG3, MDM2, MYCN, NCOR2, PRDM1, PTPRB, RFWD3, SETD1B, ZCCHC8 and ZFHX3. From the amplified genes identified 35 were recurrently affected (Figure 14b) and we find these genes to be significantly enriched for COSMIC cancer consensus genes (P = 0.002, CI of odds ratio 2.12-∞, one-sided Fisher's exact test). An enrichment was still evident after removing MYCN from the list (P = 0.009, CI of odds ratio 1.61-∞, one-sided Fisher's exact test). However, we did not detect an enrichment of COSMIC census genes in the complete list of amplified genes. Furthermore, we find MYCN, DDX1, NBAS, LRATD1 (FAM84A), CYRIA (FAM49A), AC011897.1 and RP11-527L4.2 to be amplified in three or more samples, of which all but RP11-527L4.2 reside on 2p24, the chromosomal region of the MYCN amplicon. We speculated that amplification of transcription factors could be an effective mechanism to deregulate a wide range of target genes in trans and defined a set of 1765 transcriptional regulators by GO annotation GO:0140110 (transcription regulator activity) that was assigned to genes in the ENSEMBL database (version 101). The following transcriptional regulators were amplified in one or more samples: ATXN7L3, BRCA1, CCDC62, CCND1, CEBPB, CNOT2, DDX1, E2F6, E2F7, ETV4, EZH1, FOSL2, GTF2H3, KLF11, LPIN1, MEOX1, MLX, MLXIP, MYCN, MYRFL, NCOR2, NFAT5, PRDM1, PSMC3IP, PSMD9, SETD8, THAP2, UBTF, WDR43, ZFHX3, ZNF19 and ZNF821. However, we did not find transcription regulators enriched in the set of amplified genes (P = 0.3131, one-sided Fisher's exact test). Additionally, we tested for enrichment of REACTOME pathways among amplified genes. After adjusting for multiple testing we find "Hydroxycarboxylic acid-binding receptors" to be the only pathway significantly enriched (FDR = 0.018, Benjamini-Hochberg adjusted one-sided Fisher's exact test). Amplified receptors from this pathway comprise HCAR1, HCAR2, HCAR3, which are all located on 12q24.31 and are found co-amplified in the single sample NBL27. We investigate selected pathways that were nominally significant but did not reach significance after adjusting for FDR. Amplifications in the pathway "Cell cycle" (P = 4.07×10^{-3}) affect 7 samples and include BRCA1 as well as recurrently amplified genes CDK4 and MDM2. Amplifications in the pathway "PTK6 Regulates Cell Cycle" (P = 5.66×10^{-3}) comprise genes CCND1 and CDK4 and affect samples NBL04, NBL27 and NBL39. Amplifications in the pathway "Stabilization of p53" (P = 2.13×10^{-2}) affect 5 samples and include the recurrently amplified gene MDM2 as well as proteasome genes PSME3, PSMD7 and PSMD9. Supplementary table 9 lists the top 30 pathways from the amplification

[10] COSMIC - Catalogue of Somatic Mutations in Cancer, version 92

enrichment statistics based on the REACTOME database, their nominal enrichment p-value and the respective FDR.

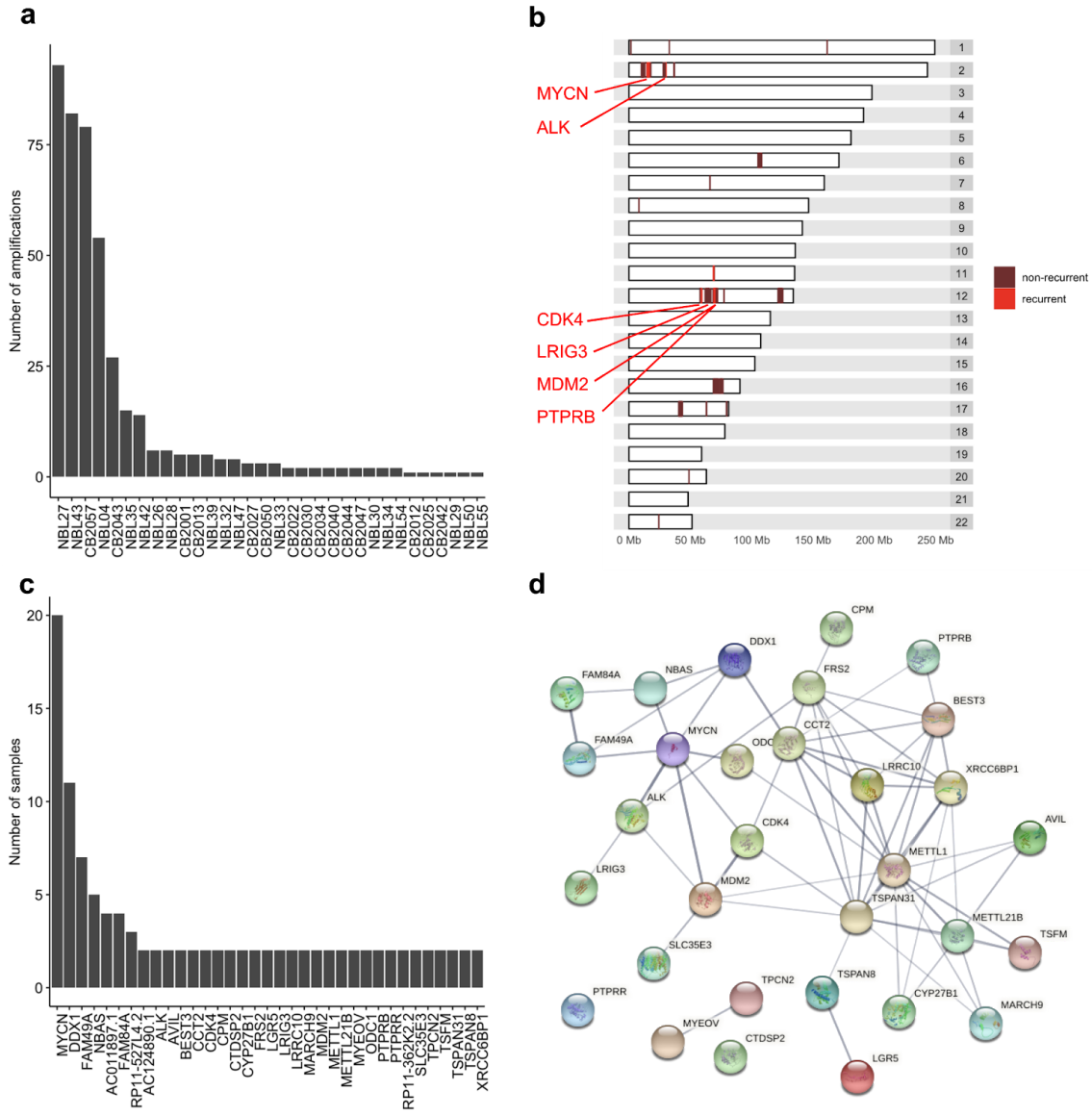


Figure 14: Amplified protein-coding genes across 116 neuroblastoma tumors. **a**, Number of amplifications per sample (samples without gene amplification are not shown). **b**, Chromosomal locations of amplified genes. Recurrently amplified COSMIC census genes are labeled by gene name. **c**, Number of samples affected by amplification in recurrently amplified genes (i.e. genes amplified in at least two samples). **d**, Network of protein interactions from the STRING database[11] for recurrently amplified genes. Thickness of edges represent data support for interactions. Network interaction enrichment P < 1.0e-16 as given by STRING network statistics.

[11] https://string-db.org/, network visualization created 1 Dec 2020

SVs were determined from the tumor and normal WGS samples. Variant calls comprise variants of classes deletion, duplication, inversion and translocation. After excluding samples without SV calls and the outlier sample NBL54, for which more than 190,000 SVs were detected, we observed between 2 and 888 SVs per sample with an average of 62 SVs. Across all SVs, we called 37% translocations, 23% inversions, 21% duplications and 18% deletions. Figure 15a shows the number of detected SVs and distribution of SV classes per sample. Generally, tumors with higher SV burden showed either almost exclusively translocations or a mix of different SV classes. More than 190,000 SVs were detected in sample NBL54 with more than 99% translocations. After excluding this high burden sample we determined the frequency of samples affected by SV breakpoints in 500 kb bins along the genome similar to the methodology used by Peifer et al. 2015. As previously reported we find the highest frequency of affected samples at the MYCN locus at chromosome 2p24 and at the TERT locus on chromosome 5p15 (Peifer et al. 2015; Valentijn et al. 2015).
Figure 15b shows the frequency of samples affected by SVs in the genomic bins on chromosomes 1-22 and X. Translocations are SVs that fuse sequences from two different chromosomes. We find multiple translocations between the MYCN locus at chromosome 2p24 and chromosome 1p and 17q. Additionally we identified multiple translocations that interconnect sequences from 11q and 17q, two chromosomes with recurrent losses and gains respectively (Figure 15c). Structural variation in a 5 Mb window around gene starts of MYCN and TERT showed breakpoints of both interchromosomal translocations and intra-chromosomal structural variants at the two loci (Figure 15d and e).

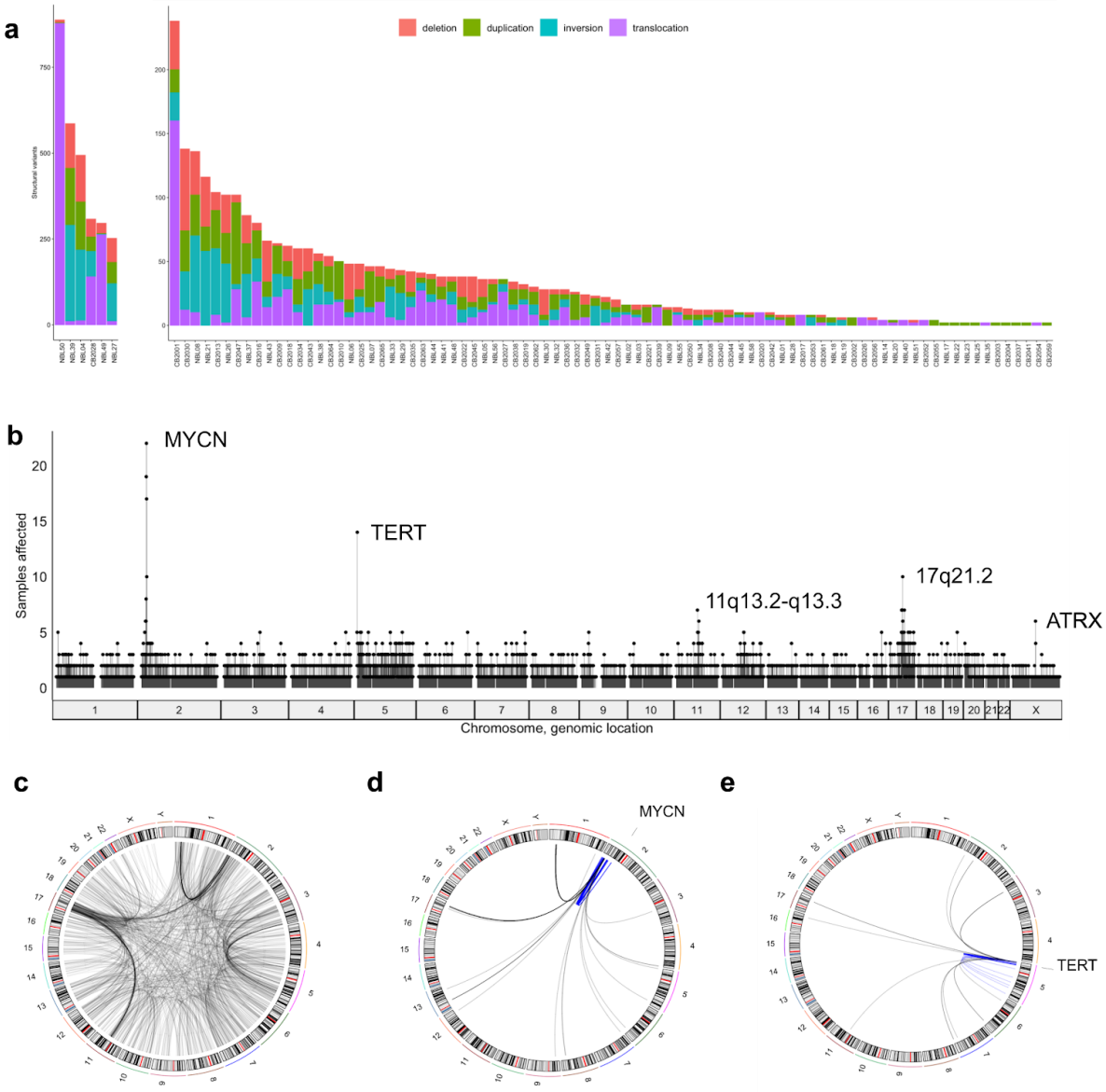


Figure 15: Detected structural variation in neuroblastoma tumors. **a**, Number of structural variants per class (deletion, duplication, inversion, translocation) and sample[12]. **b**, Number of selected samples affected by structural variation in genomic bins of 500 kb. Bins overlapping MYCN and TERT are labeled by the respective gene name. (**c**) Interchromosomal translocations. **d-e**, Structural variation with breakpoints in a 5 Mb window around MYCN (d) and TERT (e) respectively. Intrachromosomal SVs in blue, others in grey.

[12] Sample NBL54 was excluded, as it was a strong outlier with in total 190,000 structural variant calls (>99% translocations).

Somatic single nucleotide variants (SNVs) were called using WGS of tumors and WGS of matched normal samples as controls. We estimated the effect of SNVs on protein coding genes using the ENSEMBL variant effect predictor (VEP). Effect predictions were summarized in classes missense, nonsense and splice. We also determined if a gene in a sample was hit by a SV. To that extent we overlapped gene coordinates with each of the two breakpoints of a SV per sample. Deletions of ATRX exons were inferred using a targeted approach that detects local coverage differences between normal and tumor WGS alignment (Section 3.1.7) and added to the set of SVs. We aggregated gene amplification status, SV hits and SNVs to obtain a list of alterations per gene. Figure 16 shows genes frequently affected by somatic alterations and the type of alteration detected per sample. We find MYCN amplification to be the somatic gene alteration with highest recurrence (17%) across the tumors analyzed and co-amplification of DDX1 in 11 of 20 MYCN-amplified cases. Other genes at the MYCN locus, such as NBAS, FAM49A, FAM48A/LRATD1 and AC011897.1 were co-amplified at lower frequencies. Out of 18 TERT rearranged cases identified, only 2 TERT rearrangements co-occurred with a MYCN amplification. ATRX was altered in 11 samples (9%) by either mis- and nonsense mutations or SVs, these alterations were mutually exclusive with MYCN amplifications and displayed high telomere length ratios. We detected ALK alterations in 8 samples (7%), including 6 missense mutations and two amplifications. We identified 9 missense or nonsense mutations in TTN, a gene coding for a large protein (34,350 amino acids) that is a key component of assembly and function of striated muscle fibres and 9 samples (8%) that were altered by SNVs in MUC16, a large protein (14,507 amino acids) involved in formation of mucous barrier. Overall, we find notably more somatic alterations affecting genes in high risk than in medium and low risk tumors (Figure 16).

Figure 16: Genes frequently affected by somatic SNVs, amplifications and structural variants[13]. Genes affected by four or more alterations in the 116 neuroblastoma tumors analyzed are shown. TERTr: TERT rearrangement.

3.2.2 Allelic expression imbalances are enriched for imprinted genes and are less prevalent in MNA tumors

To quantify ASE in the NB tumors we genotyped donors at predefined SNPs positions reported to have MAF > 1% in the 1000 Genomes project (Section 3.1.4). We used heterozygous SNPs overlapping expressed exons of protein coding genes (ASE SNPs) and statistical phasing to determine haplotype counts from RNA-seq (Section 3.1.5). Samples for which at least one ASE SNP was available in a given gene are considered informative for

[13] Sample NBL54 was excluded from this figure, as it was a strong outlier with in total 190,000 structural variant calls (>99% translocations).

ASE in that gene. In contrast, a sample was considered uninformative for ASE in a specific gene if no ASE SNP was detected. A single ASE SNP was identified in 37% of gene and sample pairs that were informative for ASE. However, for the majority (63%) of ASE informative gene-sample pairs multiple ASE SNPs were detected and in these instances allelic counts from the maternal and paternal allele were aggregated across the gene based on statistical phasing of the ASE SNPs (Section 3.1.4). For less than 13% of ASE informative genes more than 5 ASE SNPs and for less than 3% more than 10 ASE SNPs were used to measure ASE. Figure 17 shows the distributions of the fraction of ASE informative genes per sample for genes harboring between 1 and 20 ASE SNPs.

We find on average 5799 protein coding genes to be informative for ASE per sample (95% CI 5654-5944). ASE counts from the two alleles were used to determine AEI, a binary indicator for statistical significant imbalance in ASE (Section 3.1.5). 22.3% (95% CI 20.3%-24.4%) of genes informative for ASE across all samples showed AEI, with strong variability in the proportion of AEI genes (between 8.6% and 43.8% per sample). Figure 18 shows the number of expressed protein coding genes, and the distribution of genes with and without AEI among genes informative for ASE for each sample.

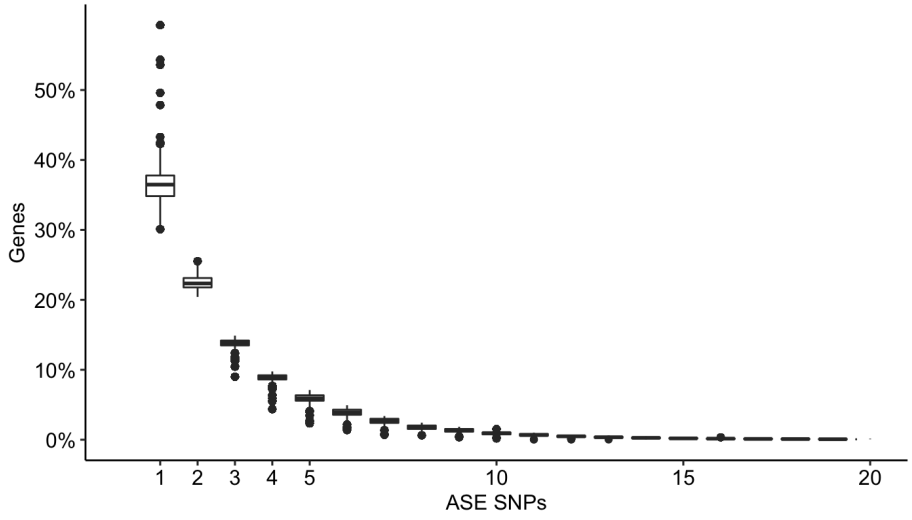


Figure 17: Distribution of fraction of informative ASE genes per sample for genes harboring between 1 and 20 ASE SNPs. Horizontal line indicates median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. Dots represent outliers.

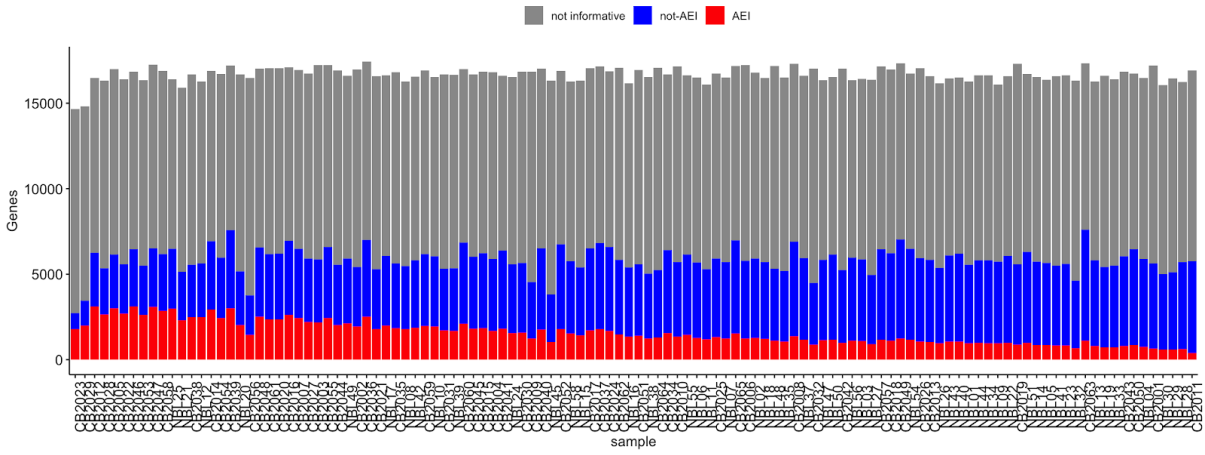


Figure 18: Number of expressed genes affected by AEI, not affected by AEI and uninformative for ASE per sample. Samples ordered by AEI frequency in ASE informative genes. Genes with a minimum of 5 variance stabilized counts were considered as expressed.

To identify genes that are currently subject to expression imbalances we summarized the AEI frequency for each gene with at least 10 ASE observations across the cohort. The AEI frequency was represented by a value between 0.0 and 1.0, where 0.0 indicates that none of the ASE informative samples showed AEI and 1.0 indicates that all ASE informative samples showed AEI in that gene. Additionally, to measure the strength of allelic expression, we determined the mean ASE ratio for all genes considered for AEI frequency. On average 24% of samples showed AEI and the average mean ASE ratio of these genes was 0.61 with a relative narrow distribution (SD 0.03). Thus, for most genes a minority but substantial proportion of samples showed AEI and across the cohort ASE of these genes was only moderately imbalanced. However, we found a subset of genes with AEI frequency above 0.75 that additionally showed extreme average ASE ratios. 44 of 45 informative samples harbored AEI in PEG10, the gene with the highest AEI frequency across the cohort that also showed the highest average ASE ratio (0.96), indicating that expression of this gene was limited to one allele in almost all samples. PEG10 is known to be imprinted on the maternal allele and expressed from the paternal allele only (Ono et al. 2001). In light of this finding we thought to identify imprinted genes among those for which we determine AEI frequencies and ASE strengths. To that extent we annotated genes by imprinting status (Morison, Ramsay, and Spencer 2005). Supplementary table 6 lists Ensembl identifiers of genes considered to be imprinted in this analysis. Besides PEG10 this revealed additional imprinted genes among those with AEI frequency > 0.75 and ASE ratios > 0.75: L3MBTL1, IGF2, SNRPN, SNURF, PEG3, RTL1, DLK1, PLAGL1, NAP1L5 and GRB10. Figure 19a shows AEI frequency, mean ASE ratio and imprinting status of all genes considered. Next,

we tested for enrichment of imprinted genes among higher AEI frequency and average ASE ratios. We find both AEI frequency and average ASE ratio to be significantly higher in imprinted genes (P = 0.003 and P = 3×10^6, one-tailed Wilcoxon rank sum test). Figure 19b and c show the distribution of AEI frequencies and mean ASE ratios of imprinted compared to other genes respectively.

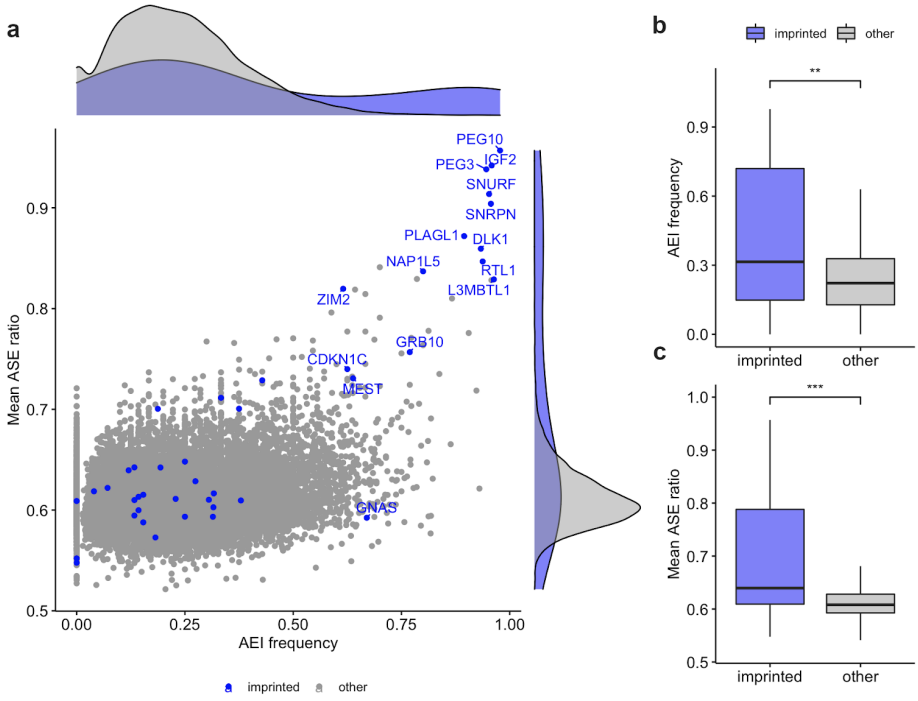


Figure 19: Comparison of genes by frequency of allelic-expression imbalance and mean allele-specific expression ratio across samples. **a**, AEI frequency and mean ASE ratio per gene. Genes with known imprinting status in blue, others in grey. **b**, AEI frequency by imprinting status per gene. **c,** Mean ASE ratio by imprinting status per gene. Midline in boxplots marks median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. ***: P < 0.001, **: P < 0.01, two-sided Wilcoxon rank sum test.

We characterized tumors by the number of protein coding genes affected by AEI and the number of protein coding genes subject to copy-number imbalances (i.e. major allele count > minor allele count). Tumors showed substantial heterogeneity both in the number genes with imbalanced expression and copy-number imbalances. Across the 116 tumors, the number of AEI ranged between 380 and 3,187 (median 1,731). CN imbalance genes ranged from 0 to 5,521 (median 1,431).

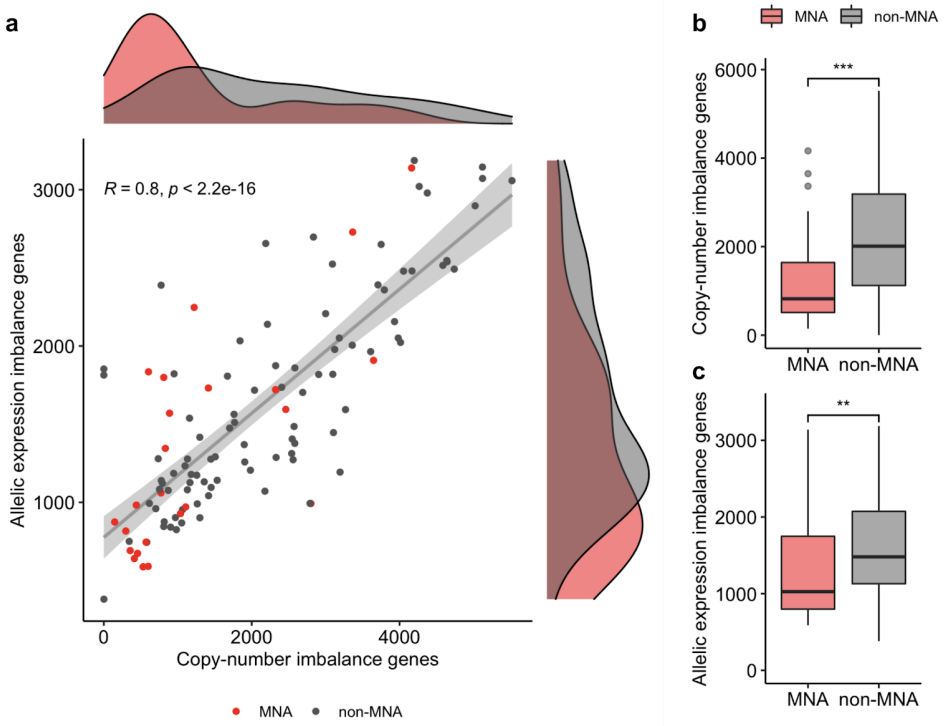


Figure 20: Comparison of samples by copy-number- and expression imbalances. **a**, Number of genes affected by copy-number imbalance vs. number of genes affected by allelic expression imbalance per sample. **b**, Number genes affected by copy-number imbalance between samples with and without MYCN-amplification (Wilcoxon rank sum test, P=0.000394). **c,** Number genes affected by allelic expression imbalance between samples with and without MYCN-amplification (Wilcoxon rank sum test, P=0.01569). MNA: MYCN amplification. Midline in boxplots marks median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. ***: P < 0.001, **: P < 0.01, one-sided Wilcoxon rank sum test.

We found a strong correlation between these two observations (R=0.8, P < 2.2 × 10^{16}). In tendency, MNA tumors showed lower numbers of genes affected by AEI and CN imbalances (Figure 20a). MNA tumors had a median of 821 CN imbalanced genes compared to 1,945 in the non-MNA group. Similarly, MNA tumors had a median of 1,026 AEI genes compared to 1,461 AEI genes in the non-MNA group. We tested for association between MNA status and these imbalances and found both the number of CN-imbalanced genes (P = 1.97 × 10^{4}) and AEI genes (P = 0.008) to be significantly lower in MNA tumors (one-sided Wilcox rank sum test). Figure 20b and c show the distributions of CN imbalance genes and AEI genes for MNA an non-MNA tumors respectively. Our findings indicate that both expression and copy-number imbalances are less prevalent in tumors that are driven by MYCN. However, we also find 6 out of 24 (25%) MNA tumors to harbor a higher number of CN imbalances

than the median non-MNA samples, suggesting that a subset of MNA tumors is characterized by a high burden of CN imbalances. Conversely, 14 out of 90 (16%) non-MNA tumors had lower burden of CN imbalances than the median of MNA tumors, out of which 7 and 7 were from the low and high risk group respectively. As gains and losses are the basis for CN imbalances, these results are in line with our earlier observations of (1) two distinct clusters of MNA tumors, that show high and low burden of overall CN alterations respectively and (2) a subset of low risk samples with low burden of CN alterations (Figure 13).

3.2.3 Somatic copy-number is a major genetic driver of expression and ASE

To determine cis-regulatory and local genetic effects that influence gene expression in the NB primary tumors we modeled ASE and total expression by germline and somatic variation detected in WGS at the respective gene loci and by additional covariates that we expected to have an influence on the measurements (Section 3.1.9). Briefly, ASE and total expression were modeled by genotypes of germline SNPs identified by eQTL and aseQTL mapping and somatic effects from copy-number variation, SV breakpoints at the gene locus (including flanking regions) and SNVs at promoter and gene coordinates. Additional covariates comprise tumor purity, MNA status and ASE SNP coverage (for ASE model only). The MNA covariate was added, because we expected upregulation of MYCN to have trans regulatory effects introducing expression variance of MYCN targets between MNA and non-MNA tumors. We required at least 20 complete sample observations per gene. 10,617 and 11,809 genes were considered for the genome-wide analysis of variance components of ASE and total expression. We estimated the total average contribution of genetic effects to ASE and total expression across all genes and samples by the median of the resulting distributions per effect. Copy-number effects showed the largest effect from all genetic factors considered and we estimated it to explain 26.2% of ASE variance and 7.8% of total expression variance. Germline cis-regulatory effects were estimated to be the second largest local genetic contributor: Together, eQTL and aseQTL explained 6.6% of variance in ASE and 3.2% of variance in total gene expression. We found that aseQTL effects explain a higher proportion of ASE variance than eQTL effects (6.0% compared to 0.5%), and conversely, in total expression eQTL effects explain a larger share of variance than aseQTL effects (2.5% compared to 0.6%). With less than 0.1% of and 1.2% somatic SVs and SNVs explain the fewest amount of variance in ASE and total expression respectively. Together, tumor purity and MNA status were estimated to contribute to 2.5% of ASE variance and 7.8% of expression variance, and ASE SNP coverage was responsible for 3.3% of ASE variance.

While MNA explained 5.6% of variability in total expression, it only contributed to 1.3% of variability in ASE. Notably, the largest amount of variation in both ASE and total expression remained unexplained by our model and the amount of unexplained total expression was considerably larger than for ASE. On average 50.3% of variance in ASE and 80.0% of variance in total gene expression were neither explained by genetic effects nor other covariates considered. Our findings show that among local genetic and cis-regulatory factors somatic copy-number variation has the largest effect on both ASE and total gene expression in neutoblastoma. Germline cis-regulatory variation is the second largest contributor and somatic SVs and SNVs display only minor effects on overall expression variability.
Figure 21a and b show the distribution of the fraction of variance explained by local genetic and cis-regulatory effects as well as additional covariates for ASE and total expression respectively.

We determined significant effects of somatic variation to ASE and expression to individual genes by the ANOVA's F-statistic (Section 3.1.9). The corresponding germline QTL effects will be investigated in greater detail in chapter 5. ASE of 8,192 and total expression of 8,580 genes were detected to be significantly affected by copy-number. We will analyze copy-number dosage effects on total expression in section 3.2.5. We find ASE in 11 genes to be associated with somatic SVs and SNVs, but we did not find recurring somatic gene hits with similar ASE ratios (Supplementary figure 1-3). 14 genes were associated with significant effects of SVs on total expression, including TERT, SLC6A18 and SLC6A19 (the latter two located downstream of TERT on 5p) as well as MYCN, DDX1 and NBAS, which are frequently co-amplified in MNA tumors (Supplementary figure 4). Notably, SV effects on genes at the MYCN locus co-occurred with MYCN copy-number amplification status of that sample. We could not examine the effect of SVs on ASE of TERT, because only two samples were informative for the phenotype for this gene. Total expression of 17 genes were significantly affected by SNVs, but similar to the related ASE results for this class of variants, we did not detect consistent regulatory effects in two or more samples for any gene (Supplementary figure 5 and Supplementary figure 6). Supplementary table 7 and Supplementary table 8 list the variance explained per covariate per gene and the covariate's p-value for based on ANOVA's F-statistic for the analysis of ASE and total expression respectively.

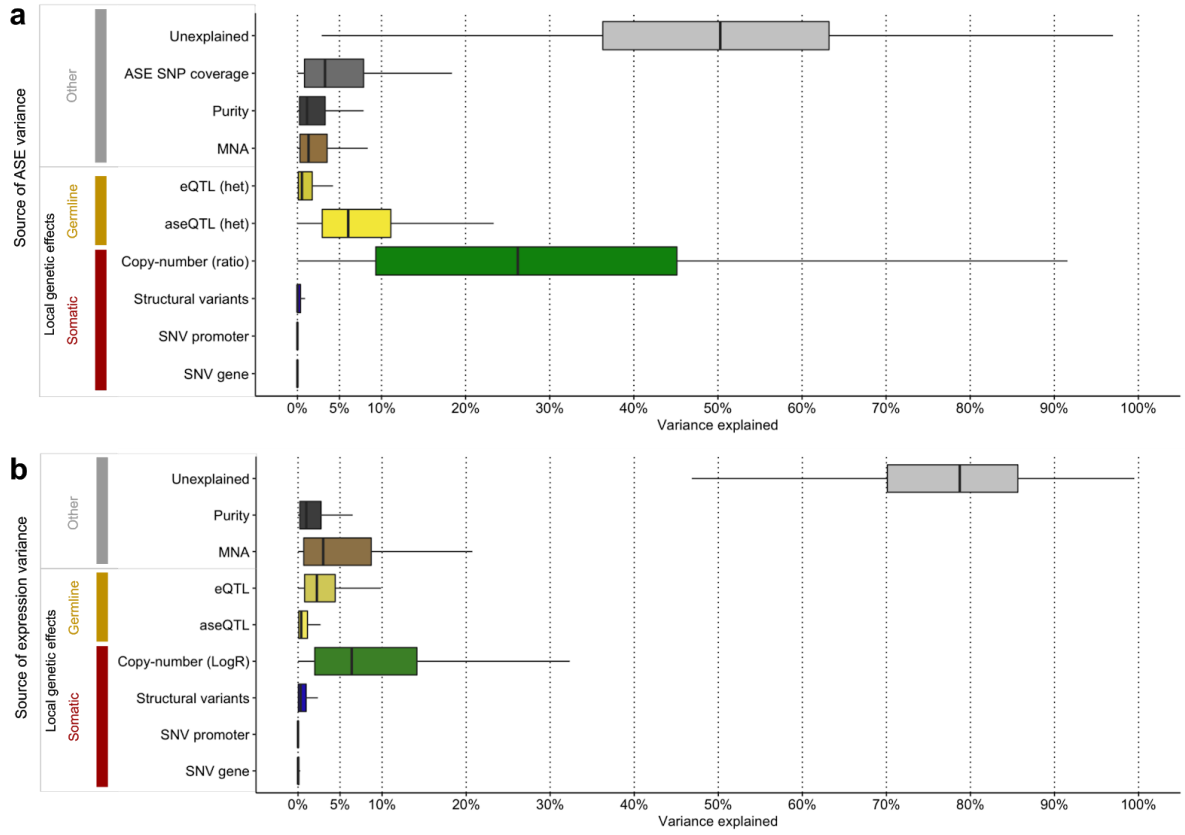


Figure 21: Quantification of local genetic effects as sources of variance. Genetic effects on (**a)** allele-specific expression (ASE) and (**b**) total gene expression. MNA: MYCN amplification, eQTL: expression quantitative trait locus, aseQTL: allele-specific expression quantitative trait locus, LogR: tumor/normal coverage log ratio, het: heterozygosity, SNV: single nucleotide variant. Midline in boxplots marks median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR.

We next investigated how somatic copy-number imbalances affect ASE and AEI genome-wide. Examination of ASCN profiles and results from the AEI test revealed that regions of chromosomal and segmental gains and losses are spatially linked to coordinates of genes with small p-values from the binomial test for AEI. As an example, Figure 22 shows the ASCN profile and AEI test p-values for genes for tumor NBL07. To quantify the effect of CN imbalances on ASE and AEI we classified genes into copy-number imbalance states *balance*, *weak imbalance*, *strong imbalance*, *amplification* and *LOH* (Section 3.1.6). We found that the majority (64.4%) of genes across all tumors displayed a balanced copy-number state, in which both major and minor allele are equally abundant. Unbalanced CN alterations introduced weak imbalances (28.8%), followed by LOH (3.9%), strong imbalances (2.8%) and focal amplifications (0.03%) (Figure 23 c). We compared the fraction of samples affected by AEI per gene to occurence of these five copy-number balance states.

This revealed that regions of frequent CN imbalances show a tendency towards elevated AEI frequencies (compare Figure 23a and b). AEI frequencies were notably increased on chromosome arms 17q, 11q, 1p and chromosome 7. And these chromosome arms also displayed high recurrence of weak and strong imbalances (17q, 7) as well as LOH (11q, 1p). Next we calculated the fraction of AEI genes for each of the CN balance states (Figure 23) and found that genes in focal amplifications were most frequently detected to harbor AEI (91.6%), followed by genes in LOH regions (72.7%), strong imbalances (69.1%) and weak imbalances (48.4%). In regions of balanced copy-number AEI was still detected in 12.7% of genes. We compared distributions of ASE ratios between copy-number balance states and found the average ASE ratio increased by the strength of imbalance, with amplifications and LOH regions showing the highest ASE ratios. The median ASE ratio in balanced CN was 0.55 compared to 0.91 in focal amplifications. The differences in ASE ratios of increasing imbalance states as well as between focal amplifications, LOH and strong imbalances were highly significant (all P < 0.001, two-sided Wilcox rank sum test). Figure 23 shows the distribution of ASE ratios per gene by CN balance state. These findings show that chromosomal CN imbalances are linked to expression imbalances and the degree of allelic imbalance is proportionally related to the degree of expression imbalance in RNA. The stronger the CN imbalance was the more AEI genes were detected. Focal amplification and LOH show both the highest frequency of AEI and strongest allelic skews in expression, while balanced CN regions show less allelic skew and fewer AEI genes.

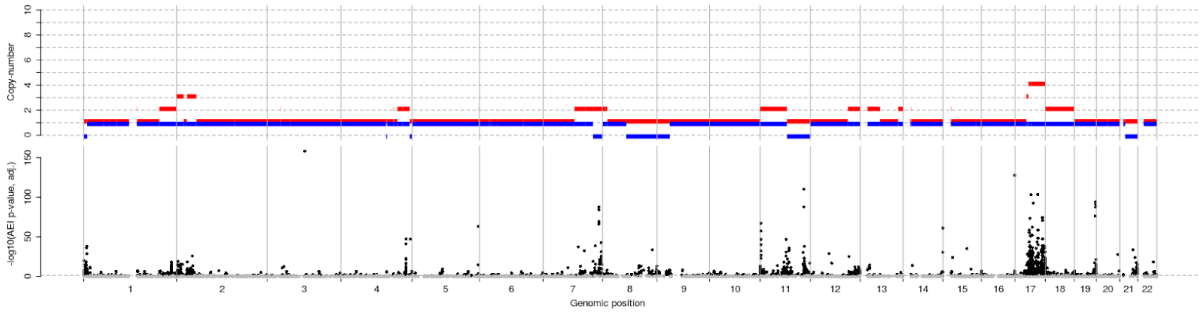


Figure 22: Genome-wide allele-specific copy-number and expression imbalances in NBL07. ASCN profile (top) and -log10 p-value of allelic expression imbalance test per gene informative for ASE (bottom) of sample NBL07. Red: major allele copy-number. Blue: minor allele copy-number. Black dots: AEI genes (AEI test FDR < 0.05, Benjamini-Hochberg). Grey dots: Genes without detectable AEI.

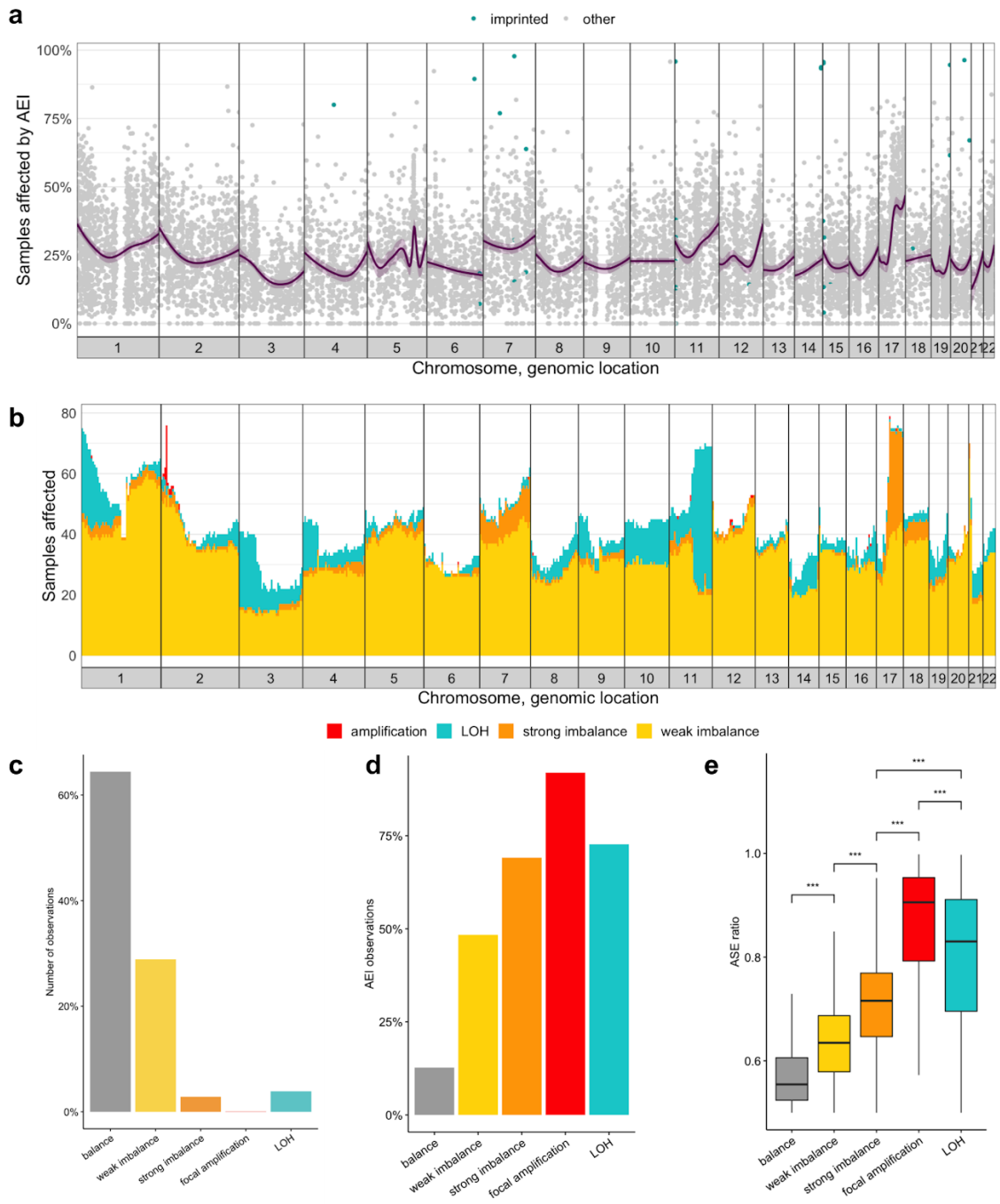


Figure 23: Genome-wide frequencies of allelic expression imbalance and somatic copy-number imbalances in 116 primary tumors. **a**, AEI frequency per gene. Dark purple line: Smoothed average AEI percentage. Light purple ribbon: 95% confidence interval of average AEI percentage. **b**, Number of samples affected by copy-number imbalances summarized in 5Mb genomic bins. **c**, Number of observations (gene-sample pair) per copy-number balance state. **d**, Percent of observations with allelic expression imbalance. **e**, Distribution of ASE ratios, outliers not shown. Midline in boxplots marks median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. AEI: allelic expression imbalance, LOH: loss of heterozygosity. ***: two-sided Wilcoxon rank sum test P < 0.001.

3.2.4 Amplified genes show strong expression from the highly abundant allele

Our copy-number analysis classified 357 genes from 32 tumors as amplified (Section 3.1.6 and 3.2.1). To investigate gene expression consequences of these extreme CN increases we analysed total expression and ASE of amplified genes in conjunction with measures obtained for allelic and total copy-number. We first focused on MYCN, the gene with the highest number of amplifications across the tumors analyzed (Figure 16) to see how DNA and RNA readouts at this locus were related and which measures would best separate MNA from non-MNA tumors as defined by the clinical annotation. We compared LogR and gene expression across all tumors and found MYCN expression to be higher in all but one tumor (CB2047) annotated as MNA in the clinical data compared to non-MNA tumors (Figure 24 a). We compared these measures by Wilcox rank sum test and found both differences in expression ($P = 1.9 \times 10^{12}$) and LogR ($P = 1.7 \times 10^{9}$) to be highly significant (Figure 24b-c). We next compared total gene expression with the CN ratio of the CN segment overlapping MYCN and found higher ratios in all but two samples (CB2047, CB2024), for one of which (CB2047) we could not detect increased expression levels of MYCN (see above). MNA samples showed significantly ($P = 2.0 \times 10^{12}$, two-sided Wilcox rank sum test) different CN ratios than non-MNA tumors (Figure 24e). We found MYCN in 32 tumors to be informative for ASE, four of which showed MNA. All those samples had extreme MYCN ASE ratio (> 0.9) and ASE ratios were significantly different between MNA and non-MNA samples ($P = 1.6\times 10^{3}$, two-sided Wilcox rank sum test) (Figure 24f). These findings show that MYCN amplification status is associated with increased LogR and expression as well as strong allelic skews in DNA as determined by ASCN analysis and RNA as determined by ASE analysis. None of the measures could perfectly separate MNA and non-MNA cases, but CN ratio in conjunction with total gene expression showed best separation. ASE provided perfect separation, but due to the reduced number of tumor samples that were informative for ASE at MYCN we cannot compare it to the other measures in that respect.

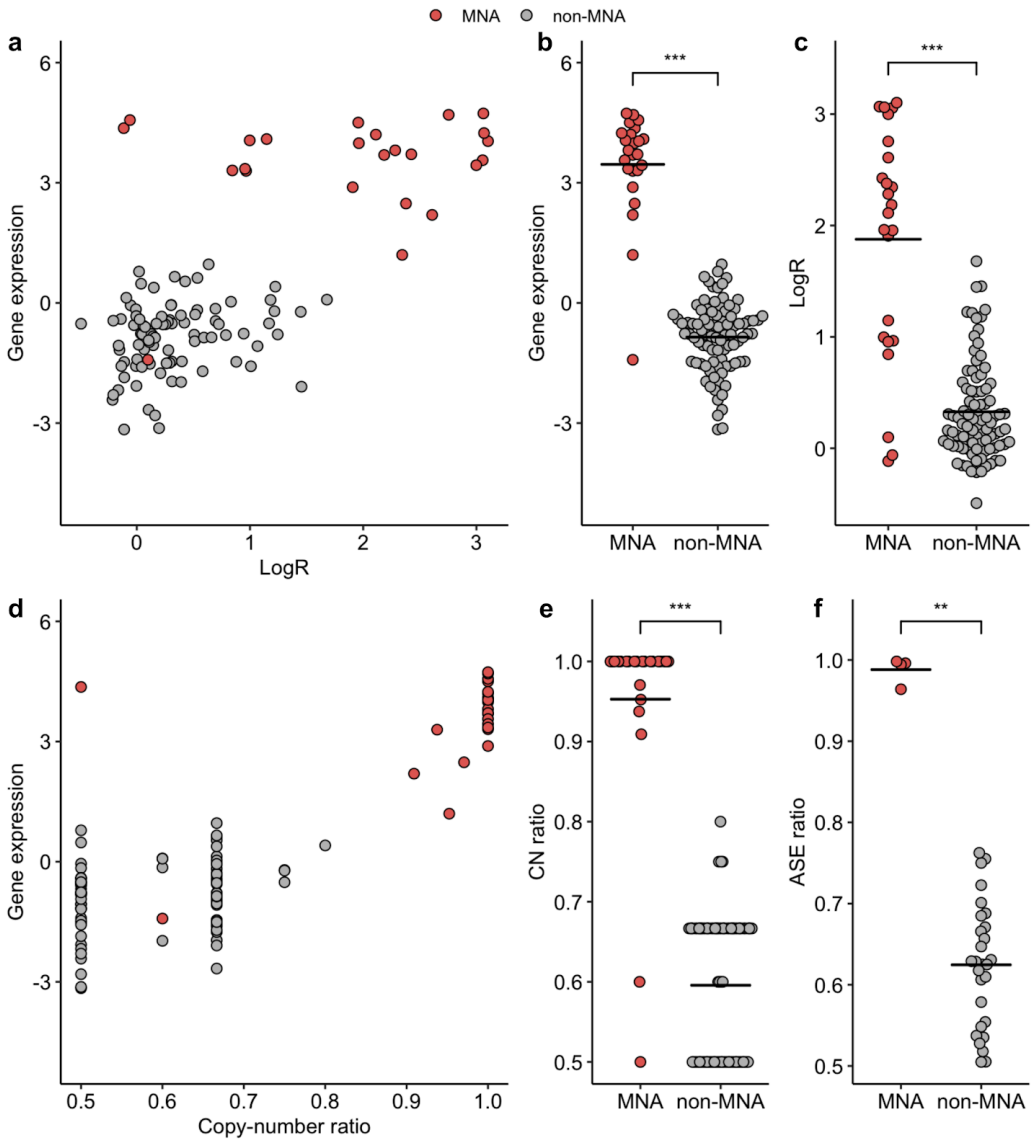


Figure 24: Copy-number, expression and allelic skews of MYCN amplifications. Each dot represents a tumor sample. **a**, MYCN expression and relative abundance of DNA reads (LogR) at MYCN gene locus. Comparison of (**b**) Gene expression and (**c**) LogR between MYCN amplified and other tumors. **d**, MYCN expression and copy-number ratio per sample. Comparison of (**e**) copy-number ratio and (**f**) ASE ratio between MYCN amplified and other tumors. MNA: MYCN amplification as defined by clinical annotations. Horizontal bar indicates mean value. Two-sided Wilcoxon rank sum test: *** P < 0.001, ** P < 0.01. Tumors without clinical MNA annotation (CB2054 and CB2055) not shown.

Our approach to detect gene amplifications is based on differences in total DNA abundance between normal and tumor WGS samples (Section 3.1.6). To understand the relation between the CN amplification state, DNA abundance as determined by LogR and ASE ratio we compared these variables across observations informative for ASE. Of 432 observations we determined as amplified, 192 were informative for ASE. We compared ASE and LogR of these observations with all other ASE informative observations and found that the majority (54.7%) of amplified genes harbored extreme ASE ratios (> 0.9). However, conversely the majority of observations with extreme ASE ratios were not classified as amplifications and even tended to show lower LogR (see trendline in Figure 25a), a fact we attributed to observations of LOH, which show similarly high ASE ratios, overall higher abundance than amplifications (Figure 23 d-e) and most frequently (79%) correspond to CN losses. Few amplifications (12.5%) showed high LogR (> 2) despite moderate ASE ratios (< 0.9). However, we found 32.8% of observations to have moderate ASE ratios (< 0.9) and no extreme LogR (< 2). Earlier, we showed that genes that overlap focal amplified CN segments had significantly higher ASE ratios compared to genes with other CN states (Figure 23e). Yet, it remains to be investigated if the allele which dominates the copy-number imbalance corresponds to the allele, that is more strongly expressed. To this end we determined the major allele RNA fraction and compared it to the CN balance state. Similar to the ASE ratio the major allele RNA fraction reflects the strength of expression imbalance, but unlike the ASE ratio it is adjusted to the CN imbalance, so that values above 0.5 correspond to an agreement of allelic skews between CN and ASE, and where values lower than 0.5 correspond to a lower expressed of the minor allele. We compared the distribution of major allele RNA fraction to CN balance states and found amplifications to harbor strongest expression skew towards the major allele (Figure 25). With a median of 91% of reads amplifications showed the strongest expression preference for the major CN, followed by LOH (92.9%) and strong imbalances (71.4%). Differences in major allele RNA fraction between amplifications and all other CN balance states were highly significant ($P < 3.30 \times 10^{14}$ or smaller, two-sided Wilcox rank sum test).

We showed that total expression of MYCN is significantly higher in MNA tumors (Figure 24b) and we seeked to understand if this is generally true across all amplifications. To this end, for each sample and gene we determined the expression percentile, which corresponds to the percent of samples that show equal or lower expression per gene. We then compared these percentiles across CN balance states and found amplifications to harbor the highest median value (98.2 percentile) and LOH the lowest (30.6 percentile). Differences in percentiles

between amplifications and all other CN balance states were highly significant ($P < 2.22 \times 10^{16}$). Supplementary table 5 lists all identified amplification candidates, their tumor/normal coverage ratio and the corresponding sample's within-gene expression rank.

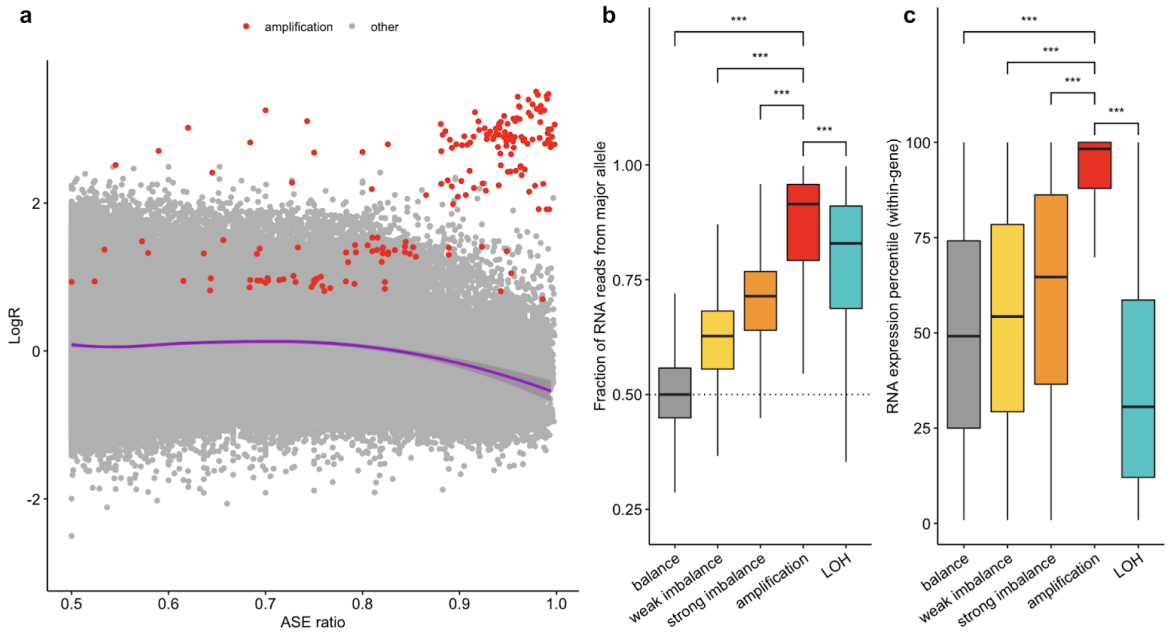


Figure 25: ASE, allelic expression preferences and expression strength of gene amplifications. **a**, Relative abundance of DNA reads and ASE ratios for observations (genes × samples) informative for ASE. Loess smoothing and 95% confidence interval in violett. **b**, Proportion of RNA reads from major allele for genes with different copy-number balance states. **c**, Distribution of within-gene expression rank as percentiles compared between copy-number balance states (100 corresponds to highest and 0 to lowest expression across samples). As MYCN was used to estimate amplification parameters, all MYCN observations were excluded. Midline in boxplots marks median. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. ***: two-sided Wilcoxon rank sum test $P < 0.001$, Amp. cand.: Amplification candidate, LOH: Loss of heterozygosity.

3.2.5 Copy-number dosage regulates expression of cell-cycle, DNA-repair and genome stability genes

In our analysis of genetic effects on variance in total gene expression we found 8,580 genes (50.8%) with significant contribution of LogR to total expression (Section 3.2.3). To quantify deregulation by somatic copy-number we determined the strength of the copy-number dosage effect estimated by the expression variance explained (r^2) of the LogR coefficient in our genetic effect model (Section 3.1.9). Dosage effect of those with significant LogR coefficient ranged from 2.4% to 71.0% with a median of 11.3%. We set out to analyze CN

dosage effects in the context of recurrent gains and losses and affected pathways. Copy-number alterations show patterns of recurrent losses and gains specific to chromosomal regions (Figure 12b). To estimate if a given gene is predominantly up- or down-regulated by CN dosage effects we characterized each gene with $n_g + n_l \geq 5$ by a CN recurrency score defined as $(n_g-n_l)/(n_g+n_l)$ where n_g and n_l are the number of tumors showing ploidy adjusted gain and loss of the gene respectively. Genes with $n_g + n_l < 5$ were assigned a score of 0. Positive CN recurrency scores correspond to higher frequencies of gains than losses and indicate recurrent positive CN dosage effects. Negative CN recurrency scores correspond to higher frequencies of losses than gains and indicate recurrent negative dosage effects. Figure 26 shows genome-wide expression variance explained and the CN recurrency score per gene. We compared distributions of explained variance by LogR between genes with positive and negative CN recurrency scores and did not find a significant difference (P = 0.56, two-sided Wilcox rank sum test). Next, we determined pathways affected by CN dosage effects using gene set enrichment analysis (GSEA) permutation testing (100,000 permutations) on genes ranked by the proportion of expression variance explained. Pathways at FDR < 0.05 (Benjamini-Hochberg) were considered significant. We found 203 reactome pathways enriched for copy-number dosage effects (Supplementary table 10). We collapsed pathways to identify independent enrichment using the corresponding method from R-package fgsea and found 33 significant and independently enriched pathways (Figure 27). Notably, we find “Cell cycle”, “DNA Repair”, “Regulation of TP53 Activity”, “MAPK6/MAPK4 signaling”, “PTEN Regulation”, “Regulation of RUNX3 expression and activity”, “NTF3 activates NTRK3 signaling” and “Regulation of MECP2 expression and activity” among these pathways (Figure 27). To exclude that the enrichment was solely based on tendencies towards higher or lower LogR values of genes across tumors we repeated the enrichment test, but this time replaced the gene-level scores by (1) the mean LogR across tumors and (2) the mean absolute LogR across tumors. At FDR 5% test (1) did not show any significant pathway enrichments and test (2) yielded 9 significant pathways at FDR 5% (Benjamini Hochberg) (Supplementary figure 7), none of which were found significant for dosage effects above. To determine if the enrichment of dosage effects in specific pathways were dominated by recurrent gains or losses, we examined the leading edge of enriched pathways. First, we defined gene sets of recurrently gained and lost genes by a CN recurrency score greater and less than zero respectively. We then conducted a one-sided binomial test (H1: greater) for enrichment of these two gene sets in the leading edge of enriched pathways using the proportion of gained and lost genes among all genes as the hypothesized probability of success. After adjusting for multiple

testing (separately for gains and losses) we did not find an enrichment of recurrent gains or losses among the leading edge of 203 dosage effect pathways at FDR 5% (Benjamini Hochberg).

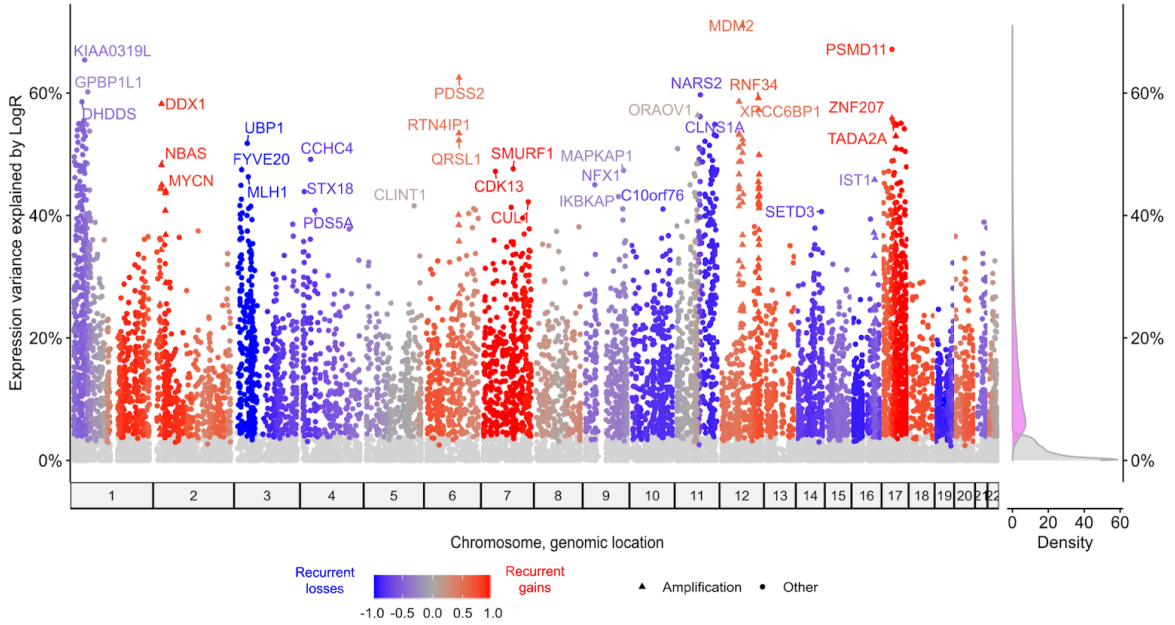


Figure 26: Genome-wide copy-number dosage effect on total gene expression. Left: Dosage effect as variance of total expression explained by LogR per gene. Dark grey and shades of red and blue indicate significant dosage copy-number effect genes (FDR < 0.05, Benjamini-Hochberg). Non-significant genes in light grey. Color scale indicates copy-number recurrency score with shades of red and blue depicting recurrent copy-number gains and losses respectively. Genes amplified in one or more tumors as triangles. Top three genes with at least 40% variance explained per chromosome are labeled. Right: Densities of percent variance explained for significant genes (magenta) and non-significant genes (light grey).

Our results show that in neuroblastoma somatic copy-number dosage affects gene expression of the majority of protein coding genes considered. We detected significant dosage effects that explain between 2.4% and 71.0% of expression variance per gene and identified equally strong effects in recurrent losses and gains. We find cell-cycle, DNA-repair and genome stability pathways to be enriched for CN dosage effects but not for trends CN gains and losses. Notably, we did not find evidence for selective pathway enrichment between recurrent losses or gains respectively.

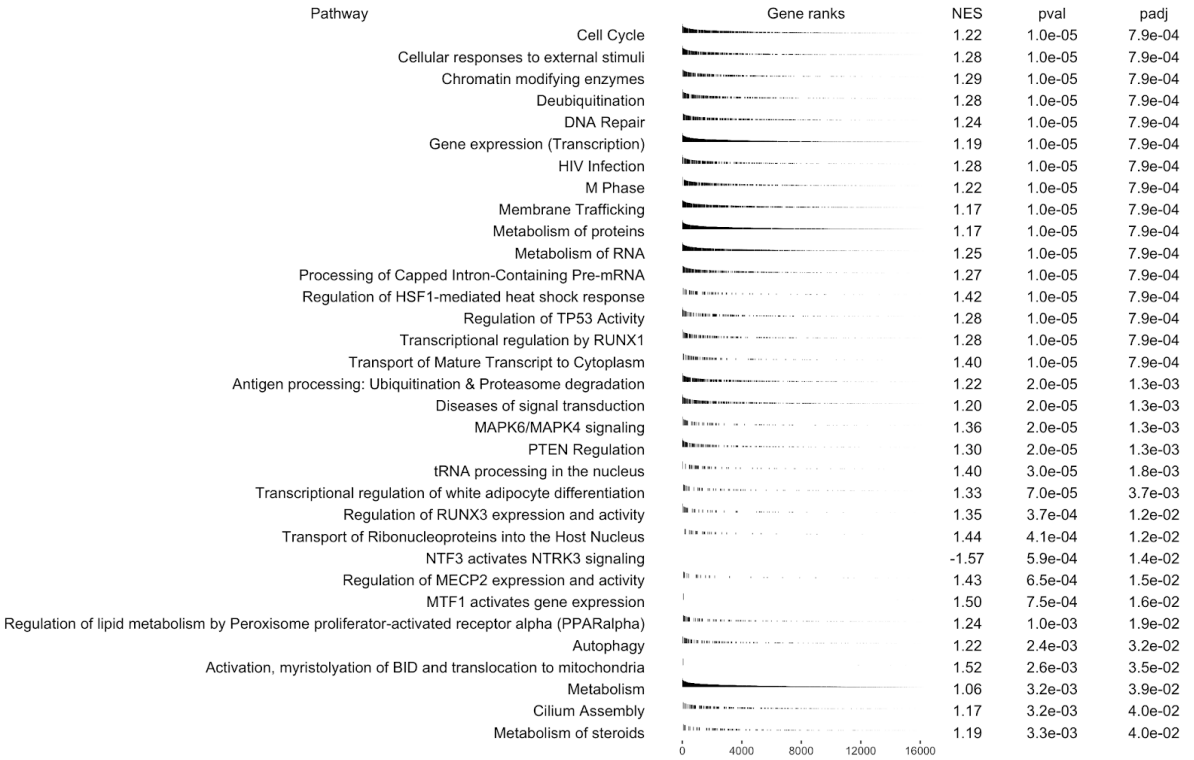


Figure 27: Independent reactome pathways enriched for copy-number dosage effect on gene expression. Black bars in column “Gene ranks” indicate pathway membership of genes ranked by copy-number dosage effect. NES: Normalized enrichment score.

3.2.6 11q loss and linked upregulation of histone genes is associated with alternative lengthening of telomeres

We found somatic CN to be a major determinant of expression variability (Section 3.2.3 and 3.2.4) and CN dosage effects to regulate genes in pathways of cell cycle maintenance and genomic stability (Section 3.2.5). CN alterations encompass large genomic regions, often on the size of chromosome arms (Section 3.2.1). In order to investigate how CN dosage-mediated gene regulation affects disease mechanisms we associated CN alterations with disease phenotypes. We aimed to better understand the molecular basis underlying the alternative lengthening of telomere (ALT) phenotype in neuroblastoma. To that extent we estimated telomere length of normal (TL_{normal}) and tumor samples (TL_{tumor}) per donor by counts of telomere repeat sequences in the respective WGS reads using the telseq method (Section 3.1.2). We used the log ratio of these two measurements, $\log(TL_{tumor}/TL_{normal})$, as a proxy for the telomere length increase in tumor tissue. Upregulation of the TERT gene is a prerequisite for canonical telomere maintenance via the telomerase pathway (Section 2.4). Comparison of TERT expression with the telomere length ratio revealed two distinct groups of high risk tumors: Those with high TERT expression and those with increased telomere

length (Figure 28a,b). Almost all tumors with high TERT expression harbored either TERT rearrangements or MYCN amplifications, confirming that these two alterations are linked to telomere maintenance by the canonical telomere pathway. We did not find MYCN amplifications in samples with long telomeres and only a single sample with both TERT rearrangement and long telomeres (NBL54). Strikingly, samples with either long telomeres or high TERT expression almost exclusively belonged to the high risk group and others predominantly to the low risk group. Investigation of ATRX alterations by targeted analysis of coverage differences between tumor and normal WGS (Section 3.1.7) and occurrence of somatic missense and nonsense mutations (SNVs) and SVs breakpoints within gene boundaries, identified 12 samples affected. ATRX altered samples had significantly longer telomeres ($P = 3.238 \times 10^{-6}$, Wilcox rank sum test). However, the majority of samples with long telomeres did not harbor ATRX alterations (see Figure 28b,c). To better understand if additional genetic traits were linked to long telomeres we tested if chromosomal copy-number differences are associated with telomere length. Based on the bimodal distribution of telomere length ratios we discriminated samples with long and short telomeres. We applied the threshold $\log(TL_{tumor}/TL_{normal}) > 0.5$ to define the ALT phenotype and then associated this phenotype with copy-number coverage differences between normal and tumor by logR averages per chromosome arm controlling for additional covariates (Section 3.1.8). We find the logR of chromosome arm 11q to be significantly associated with the ALT phenotype ($P = 1.4 \times 10^{-4}$, ANOVA Chi-squared test). Figure 28e shows results of logR and ALT association tests per chromosome arm and Supplementary table 12 lists corresponding p-values.

We further aimed to identify differentially expressed genes between samples with and without ALT using linear modeling of residual expression explained by the ALT phenotype and controlling for the same covariates used in the logR association test above. p-values were determined by a two-sided t-test of the ALT coefficient and adjusted for multiple testing burden by the Benjamini Hochberg method. At 5% FDR we find 293 differentially expressed genes, of which 143 and 150 were up- and down-regulated respectively. We find genes CCDC90B, PPME1 and NCAM1 located on 11q and RAC1 (7p) with smallest p-values among down-regulated genes. Among upregulated genes with smallest p-values we find the two histone genes H3F3B (17q), H2AFJ (12p) as well as LRRC15 (3q). Figure 29a shows p-values and ALT coefficient estimates per gene. To examine if 11q copy-number dosage effects are associated with differentially expressed genes we correlated gene expression residuals with 11q logR (Figure 29b) and found expression values of 11q genes, such as

CCDC90B, PPME1 and NCAM1 to be positively correlated with 11q logR, indicating that these genes are subject to copy-number dosage-dependent down-regulation. Interestingly, we find expression of upregulated histone genes H3F3B and H2AFJ to be substantially negatively correlated with 11q logR, as higher expression of these genes was associated with lower 11q logR (Figure 29b,d), an effect that we also observed for LRRC15 to lesser extend (Figure 29f). ALT-associated down-regulation of RAC1 was not correlated with 11q logR. However, RAC1 expression showed a substantial dosage effect by its local copy-number as determined by a significant correlation between expression residual and gene-level logR (R = 0.46, 95% CI 0.23–0.59, P = 2.722e-07). Notably, RAC1 is located on 7p and this chromosome arm's logR showed a nominal significant (P = 0.008) association with ALT, despite not reaching significance according to the 5% FDR cutoff (Figure 28e). Supplementary table 15 lists differential expression p-values, between tumors with and without ALT phenotype and correlation of expression with 11q logR per gene.

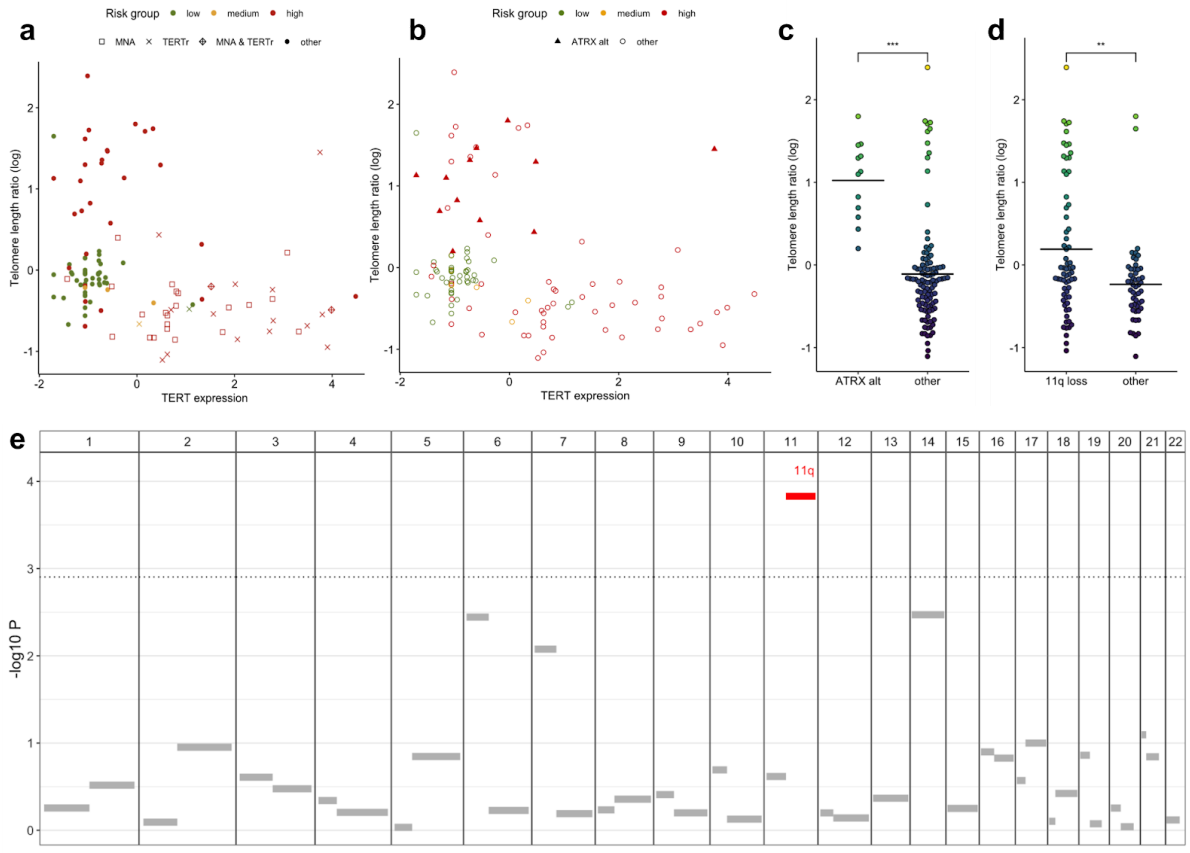


Figure 28: Chromosome 11q logR association with telomere length. Telomere length and TERT expression across samples indicating MYCN-amplification and TERT-rearrangement (**a**) and ATRX alteration (**b**) per sample. **c**, Telomere length per sample by status of ATRX alteration. ***: P < 0.001, **: P < 0.01, Wilcox rank sum test. **d**, Telomere length per sample by status of 11q loss. Horizontal bar in (c,d) indicates mean. **e**, Association results of telomere length and coverage logR per chromosome arm. Significant observations in red, others in grey. Significance threshold (FWER 0.05) demarcated by grey dotted line.

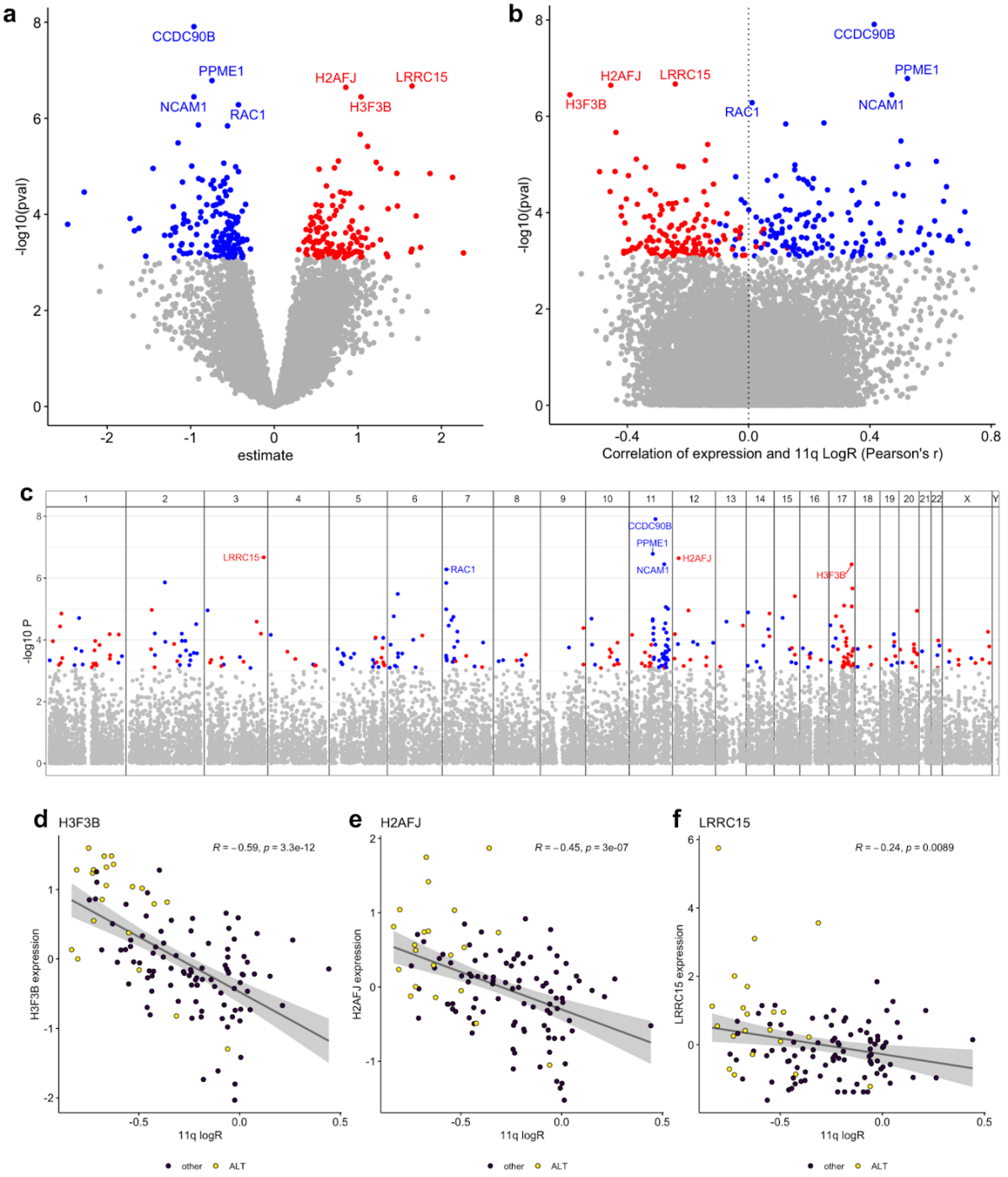


Figure 29: Differentially expressed genes between samples with long and short telomeres and their correlation with 11q logR. (**a**) p-values and effect size of differentially expressed genes. (**b**) Differential expression (telomere length) p-value and Pearson correlation between gene expression and 11q logR. (**c**) Differential expression p-value and genomic location of genes tested. Up-regulated genes in red, down-regulated in blue. Genes with -log(p-value) > 6 in (a-c) labeled by name. (**d-f**) Gene expression residual and 11q logR across tumors for differentially expressed genes H3F3B (17q), H2AFJ (12p) and LRRC15 (3q). Yellow dots indicate samples harboring alternative lengthening of telomeres phenotype.

To identify genetic effects of 11q loss on gene expression of H3F3B, H2AFJ, LRRC15, NCAM1, CCDC90B, PPME1 and RAC1 we determined ASE of these genes in our cohort. LRRC15 was only informative for ASE in four samples and the gene was not further considered. We find ASE of 11q genes NCAM1, CCDC90B and PPME1 to be significantly higher in samples harboring 11q loss and particularly strong in instances of 11q LOH, which also show lowest expression of these genes (Figure 30a top). In contrast, no ASE effect of 11q loss was apparent for histone genes H3F3B and H2AFJ and RAC1 in samples with 11q loss (Figure 30a bottom). These findings suggest that 11q-loss linked downregulation of NCAM1, CCDC90B and PPME1 in tumors with long telomeres is induced by a direct dosage-effect of the allelic loss. And upregulation of H2AFJ and H3F3B is likely due to trans-regulatory factors inducing similar upregulation from both the paternal and maternal allele of these genes on chromosome arms 12p and 17q respectively. We did not find 11q loss to be associated with ASE of RAC1, which was expected, because its expression was not correlated with 11q logR in the first place (Figure 29b and Figure 30a). However, because RAC1 is significantly affected by the copy-number dosage effect of its chromosome arm 7p, this finding provides additional evidence for a potential 7p effect on ALT that is independent from 11q loss.

Notably, despite ATRX alterations being significantly associated with longer telomeres (Figure 28c), we do not find ATRX to be differentially expressed in ALT tumors. We speculated that interaction partners of ATRX could be subject to deregulation. To identify potential interactions of ATRX with differentially expressed (ALT) genes on the level of their protein products we queried first and second degree protein interactions between ATRX and differentially expressed genes in the STRING database[14]. Interestingly, this revealed high confidence first-order interactions between ATRX and H3F3B and second-degree interactions between histone genes H3F3B, H2AFJ and H3F3C. Furthermore, we identified three first-order interactions between ATRX and differentially expressed genes PMS2, EIF2AK1 and SRSF1. Second-degree interaction involved SPIN1, which is predicted to interact with histone genes H3F3B and H3F3C, as well as second-degree interactions with IGF2BP1, ALYRED, RBFOX1 and RBFOX2, which interact with SRSF1. Figure 30b shows first and second degree protein interactions between ATRX and differentially expressed genes and indicates up- or down-regulation of these genes in ALT tumors. The observed interactions between histone genes prompted us to investigate if there was a cooperative effect of identified histone genes on the ALT phenotype. Using a generalized linear model

[14] https://string-db.org/, network visualization created 22 Dec 2020

with ALT as response and expression residuals of H3F3B, H3F3C and H2AFJ as explanatory variables we find find both H3F3B and H2AFJ to have significant coefficients (P = 0.015 and P=0.002 for H3F3B and H2AFJ respectively) and H3F3C to be not significant in the presence of the two other genes (all p-values from two-sided Student's t-test). Figure 30c depicts combinations of H3F3B and H2AFJ expression and indicates the ALT phenotype per sample.

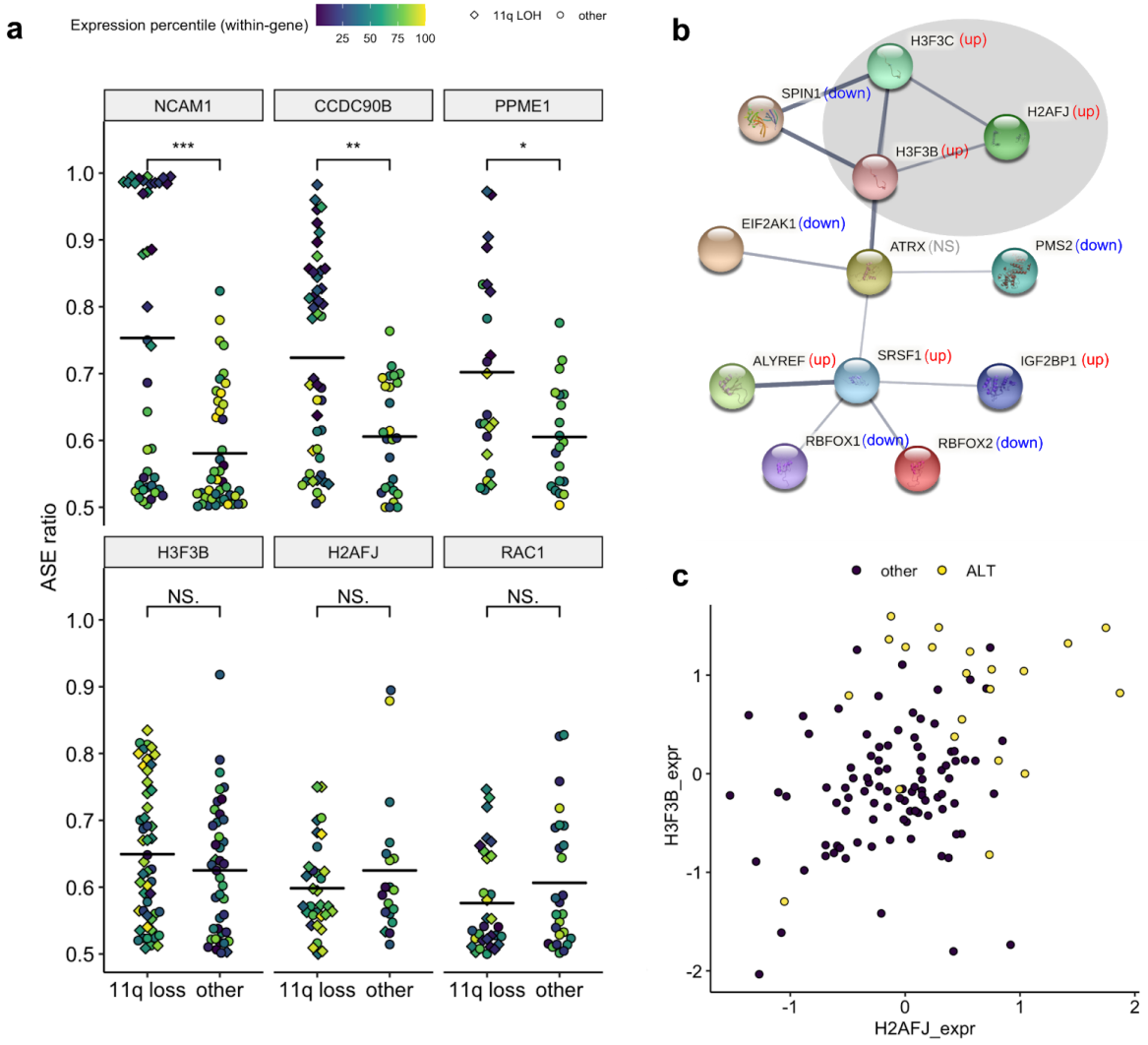


Figure 30: Local genetic and potential trans regulatory effects of 11q loss and ATRX protein interactions of ALT differentially expressed genes. **a**, Differences in ASE of 11q genes NCAM1, CCDC90B, PPME1 and distal genes H3F3B, H2AFJ and RAC1 between tumors with and without 11q loss. Diamonds indicate samples with 11q LOH. Horizontal bars indicate mean. **b**, STRING database protein interaction network of ATRX and its first and second degree interactors among differentially expressed genes (ALT). Thickness of edges represent data support for interactions. Clade of histone genes highlighted in grey. up: upregulation, down: downregulation, NS: not significant. **c**, H3F3B and H2AFJ expression residual and alternative lengthening of telomeres phenotype (indicated by yellow dots) per sample.

3.2.7 Somatic copy-number gains cooperate with TERT activation

Our results from section 3.2.3 confirmed previous findings on the role of somatic SVs in the activation of telomerase reverse transcriptase (TERT) expression, which showed that hijacking of enhancer elements to the promoter region of TERT by genomic rearrangements (Peifer et al. 2015; Valentijn et al. 2015) and increased MYCN expression (Mac, D'Cunha, and Farnham 2000; Peifer et al. 2015) are associated with TERT upregulation. We seeked to understand the interplay between MNA, TERT rearrangements (TERTr) and TERT copy-number. First, we compared TERT expression between samples with TERTr and found both MNA and TERTr samples to have significantly higher TERT expression than others (Figure 31a). We then analyzed TERT expression and tumor DNA coverage (LogR) at the locus across tumors and noted that an increased LogR was found in samples with both high and low TERT expression. We found the strongest expression of TERT among samples with higher tumor DNA coverage (LogR > 0.5), indicating a relation between copy-number gains and higher RNA levels of TERT. We also found increased coverage (LogR > 0.5) in samples with lower TERT levels, suggesting that copy-number increases may not be sufficient to elevate TERT expression in these tumors. However, samples with no apparent CN dosage effect did neither show TERT rearrangements nor MNA. We used MNA and TERTr status to stratify tumors into TERT activated and non-activated and found a strong and significant correlation ($R = 0.8$, $P = 4.8 \times 10^{-7}$) between LogR and TERT expression in TERT activated, but not in other samples ($R = 0.2$, $P = 0.08$). Interestingly we found the effect not only in MNA but also in TERTr tumors, which may point to selective gains of alleles harboring activating SVs. Figure 31b shows LogR and expression at the TERT locus for each primary tumor and results of the regression of TERT activated and other tumors respectively. To exclude the possibility that the observed effect is solely based on purity differences in TERT activated samples, we compared interaction terms between TERT activation and LogR as well as TERT activation and purity in two linear models. Linear model (1), $t \sim a + aL$, estimates TERT expression t by TERT activation coefficient a (where $a = 1$ if MNA or TERTr and $a = 0$ otherwise) and the interaction term aL, where L is the LogR. Linear model (2), $t \sim a + ap$, estimates TERT expression t by TERT activation coefficient a and the interaction term ap, where p is the estimated tumor purity. We find the interaction between TERT activation and LogR (term aL) in model (1) to be highly significant ($P = 1.12 \times 10^{-9}$, ANOVA F-statistic) and to explain 14.7% of variance in TERT expression. In contrast, interaction between TERT activation and purity (term ap) in model (2) was not significant ($P=0.196$, ANOVA F-statistic) and was estimated to explain only 0.8% of variance. Our findings show

that somatic copy-number gains can increase TERT expression, but that this dosage effect is limited to tumors in which TERT is activated by either genomic rearrangements or amplification of MYCN. It also indicates that highest TERT expression levels are found in tumors with both TERT activation and copy-number gains. In these tumors copy-number dosage effects cooperate with TERT activation to upregulate this cancer hallmark gene.

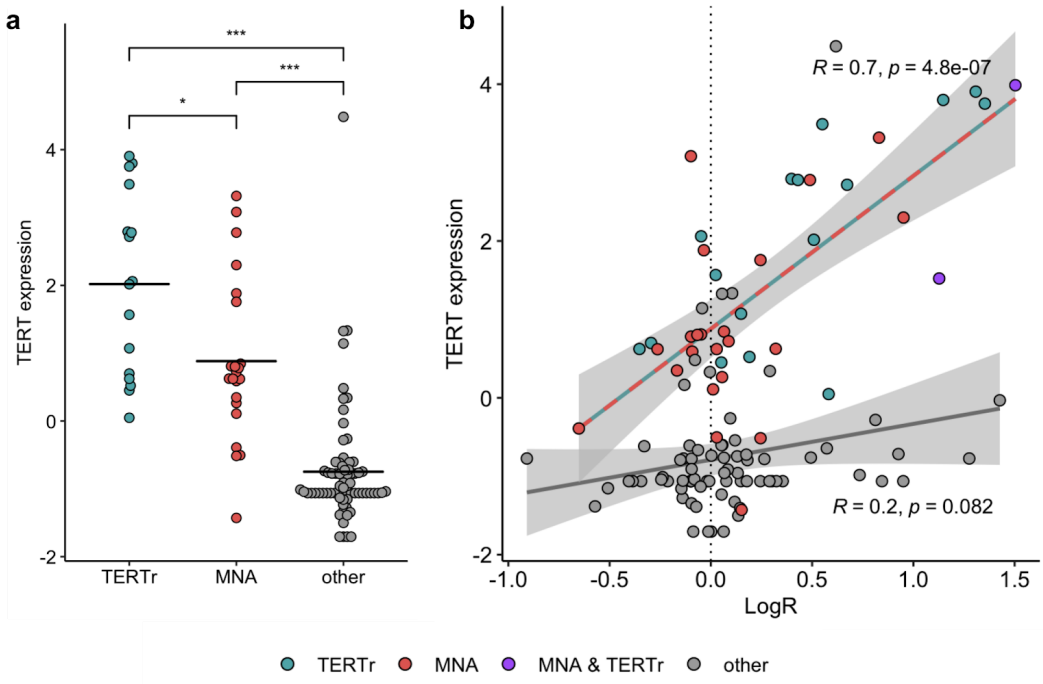


Figure 31: Cooperative effect of TERT activation and copy-number on TERT expression. **a**, TERT expression of samples harboring TERT rearrangements (TERTr), MYCN amplifications (MNA) and others. Two samples with both MYCN amplification and TERT rearrangement not shown. Horizontal bar indicates mean value. Two-sided Wilcoxon rank sum test: *** P < 0.001, * P < 0.05. **b**, Copy-number LogR and expression of TERT per sample. Regression line of samples with TERT rearrangement or MYCN amplification (TERT activated samples) striped in red and turquoise, regression line of other samples in grey. Grey ribbons indicate 95% confidence intervals for each regression line.

3.2.8 Allelic regulation associated with expression differences in survival associated genes

We aimed to identify cis regulatory effects on expression of survival associated genes. ASE is considered to be a sensitive indicator of local genetic and cis-regulatory effects and we therefore seeked to identify expression differences associated with differences in allelic regulation. To that end we correlated total gene expression with ASE ratios between all ASE-informative samples per gene. We hypothesized that this approach identifies genes for

which expression is controlled by up- or downregulation of individual alleles with differences between tumors. Because copy-number ratios showed a remarked effect on ASE (Section 3.2.3), we specifically considered its influence on ASE to distinguish copy-number dependent and independent regulation. To quantify correlation between ASE ratio, copy-number ratio and gene expression we calculated r^2 values between these three measures for each gene. Generally, we found ASE ratio and gene expression to be only weakly correlated (mean r^2 = 0.039, 95% CI 0.038-0.04). To identify gene-specific correlations, we fit a linear model for each gene with at least 20 informative samples for ASE. Using the ASE ratio as response we used the expression residual as an explanatory variable and cohort membership, tumor purity and coverage at ASE SNPs and DNA ratio as additional coefficients. The ASE–expression effect was defined as $r^2_{expr} \times \text{sign}(b_{expr})$ where r^2_{expr} is the partial r^2 of the expression coefficient and b_{expr} the corresponding model coefficient estimate. A positive effect size indicates that high ASE ratios of a gene correspond to high expression levels, whereas a negative effect size indicates that high ASE ratios of a gene correspond to lower total expression levels. We tested 10,886 genes, obtained a p-value from the ANOVA F-statistic on the gene expression coefficient and identified 467 with FDR < 0.05 (Benjamini Hochberg) as allelic regulated (AR) genes. Positive and negative ASE–expression effects as well as different degrees of CN contribution (r^2_{CN}) were found. Figure 32 shows examples of genes with strong and weak copy-number effects for positive as well as negative ASE–expression effects. Supplementary table 16 lists effect sizes and p-values of the AR test per gene.

To identify candidate genes, in which allelic regulation is associated with survival-associated deregulation, we intersected AR genes with genes differentially expressed between tumors of deceased and other patients (Section 3.1.3). Supplementary table 13 lists results of the differential expression analysis. Intersecting AR genes with 2,550 differentially expressed genes resulted in 122 differentially expressed AR genes. Figure 33a shows log2-fold change and ASE–expression effect of analyzed genes. We selected 20 genes with strong differential expression (log2 fold-change > 0.5) and substantial absolute ASE–expression effect (> 0.2) for further investigation. Notably, among this set of genes we find MYCN-amplicon genes MYCN, NBAS and DDX1. MYCN-amplicon genes were found upregulated in deceased patients by differential expression analysis and showed a positive ASE–expression effect, indicating that tumors with high ASE ratios tended to have higher expression of the respective gene. Similarly, among the selected genes we find positive ASE–expression effects for TSFM, XRCC6BP1, METTL21B, CCT2, TSPAN31, CDK4, PHB and MDM2. All

these genes except PHB were found to be amplified in at least one tumor of our cohort. Among our selected genes we identified negative ASE–expression effects for CLCN6, COL9A2, H6PD, KIF1B, RNF19B, ZNF436, PINK1, CDRT4 and RTL1. High ASE ratios in these genes are found in tumors with lower total gene expression (Figure 33b). Notably, 7 out 9 (77%) selected differentially expressed AR genes with negative ASE–expression effect were located on chromosome 1p. Among chromosomal locations of all 86 differentially expressed AR genes with negative ASE–expression effect chromosome arm 1p (56%) and 17p (12%) were most frequent, indicating that losses of 1p and 17p might underlie downregulation of these genes in tumors of deceased patients. In contrast to other genes with negative ASE–expression effect, RTL1 showed lower ASE ratios in samples with higher expression. And strong upregulation (1.60 log2 fold-change, $P = 9.9 \times 10^{-5}$) of RTL1 was found in tumors of deceased patients.

We determined variance components for genetic effects on ASE ratios among informative genes (Section 3.1.9 and 3.2.3) and investigated genetic effects on ASE for the selected AR genes specifically. We find substantial contribution (> 25% variance explained) for copy-number ratios among the majority of selected AR genes (Figure 34a). However, we identified lower CN effects for COL9A2, CDRT4 and RTL1. These genes showed comparably high contribution from germline variation identified by eQTL and aseQTL mapping, but the majority of ASE variance remained unexplained. Next, we compared ASE and CN ratios of selected AR genes between tumors of deceased and other patients (Figure 34b). Here, deceased patients were defined as those with status “deceased from disease” in the clinical annotation. Only CLCN6, DDX1, RTL1 showed nominally significant ($P < 0.05$, two-sided Wilcox rank sum test) difference in ASE ratios between patient survival. We were not able to detect a significant association between MYCN ASE ratio and survival ($P = 0.072$), likely because of missing power due to a lack of ASE informative samples for MYCN. Similarly, we compared copy-number imbalance between these two survival states and found CN ratios of CDRT4, MYCN and PHB to be significantly associated with disease-specific survival (Figure 34c).

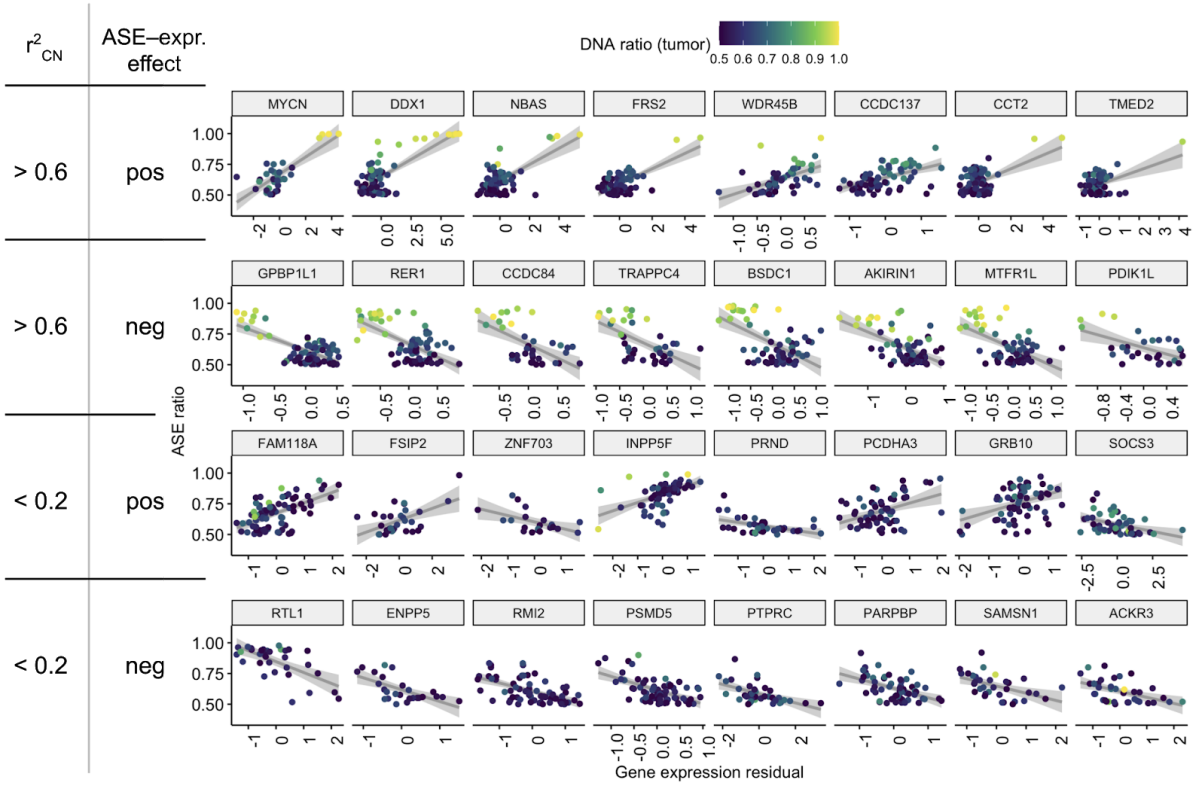


Figure 32: Gene expression and ASE ratio for AR genes with strong and weak copy-number effects per tumor. Each row corresponds to a combination of r^2_{CN} and direction of ASE–expression effect shown on the left side. The top eight genes with strongest correlation between ASE and expression are shown for each combination. r^2_{CN}: Variance in ASE explained by copy-number effect. ASE-expr. effect: ASE–expression effect.

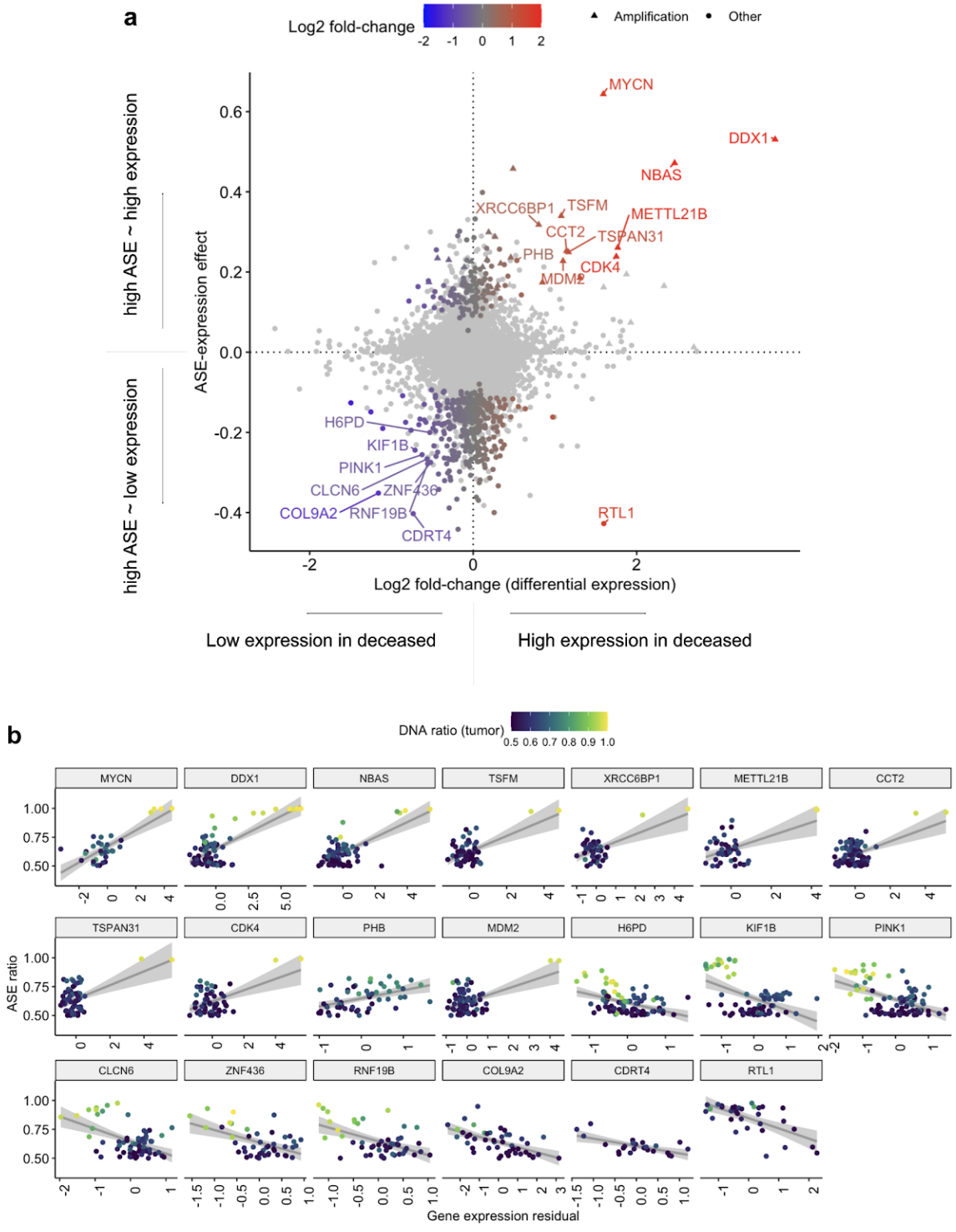


Figure 33: Differentially expressed AR genes. **a**, ASE-expression effect and Log2 fold-change of differential expression (survival) analysis. Color scale is applied to significant AR genes (FDR < 0.05, Benjamini-Hochberg), genes not significant for ASE-expression effect in light grey. Differentially expressed (FDR < 0.05 test of DEseq2 test, Benjamini Hochberg) AR genes with ASE-expression effect > 0.2 and Log2 fold-change > 0.5 are annotated by name. Genes amplified in at least one sample as triangles. **b**, ASE ratio and gene expression for genes highlighted in (a) per tumor.

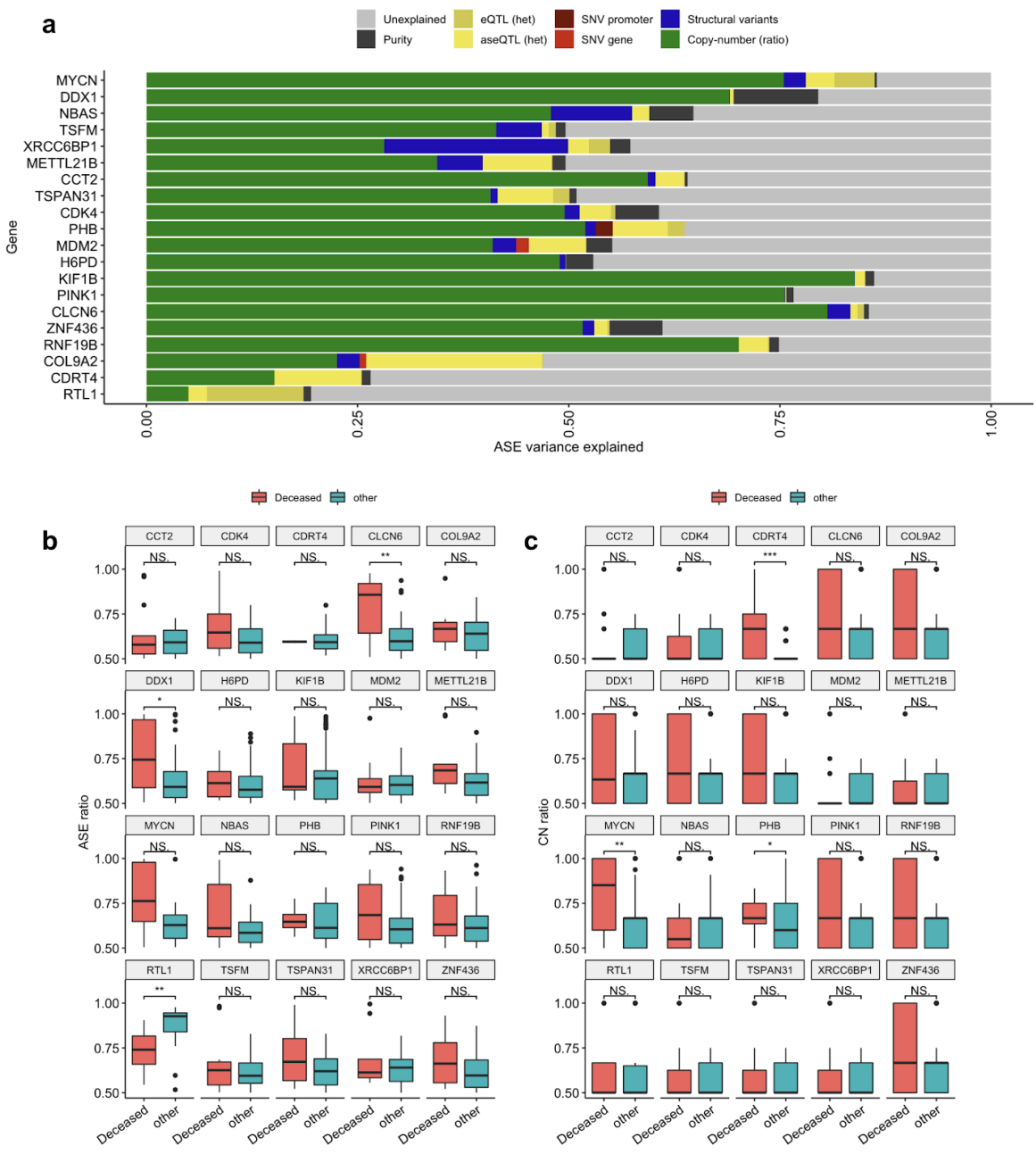


Figure 34: Variance components and survival by allelic ratios in selected allelic regulated genes. (**a**) Variance in ASE explained by genetic effects (**b**) Distribution of ASE ratios of informative observations by disease-specific survival (**c**) Distribution of copy-number ratios of informative observations by disease-specific survival. Upper and lower hinges mark the first and third quartile. Upper and lower whiskers extend to the smallest and largest value max. 1.5 × IQR. ***: P < 0.001, **: P < 0.01, *: P < 0.05, NS: not significant, two-sided Wilcoxon rank sum test.

3.2.9 17p copy-number imbalance is associated with disease-specific mortality

Our allele-specific analysis resolves genetic imbalances on both DNA and RNA in neuroblastoma tumors. We aimed to investigate if these genetic imbalances are associated with clinical phenotypes. To that end we associated somatic copy-number imbalance to disease-specific survival by generalized linear regression analysis, controlling for MYCN amplification, age, sex, stage 4 status, tumor purity and tumor ploidy. Copy-number ratios were defined based on allelic counts of major and minor allele as determined by allele-specific copy-number analysis (Section 3.1.6). In a first analysis, copy-number ratios were averaged at the level of chromosome arms using all overlapping CN segments weighted by the length of overlap with the chromosome arm and p-values were determined by comparison between a control and test model (Section 3.1.8). After multiple testing correction we found the CN ratio of chromosome arm 17p to be significantly associated (FWER < 0.05, Bonferroni) with disease-specific survival (Figure 35a). To investigate if smaller, segmental CN changes were associated with disease-specific survival we conducted a second association test on smaller chromosomal regions. Here, we summarized the CN ratio in 5 Mb bins along the genome and found four consecutive bins on 17p spanning hg19/GRCh37 coordinates 17:1–20,000,000 to be significant (FWER < 0.05, Bonferroni). Thus, after splitting chromosome 17p into smaller bins, the majority of bin on this chromosome arm reached genome-wide significance (Figure 35b). We repeated the 5 Mb association test without controlling for MYCN amplification status. Using this approach, we confirm significant (FWER < 0.05, Bonferroni) associations of CN ratio with disease-specific survival at a single bin overlapping with the MYCN oncogene (2:15000001–20000000), all bins on chromosome 17p as well as two consecutive bins on chromosome 1p at coordinates 1:55000001–65000000 (all coordinates in hg19/GRCh37). Figure 35c shows genome-wide association results of 5 Mb bins without controlling for MYCN amplification status.

We examined CN ratios, logR and states of CN segments overlapping significantly associated bins from the 5 Mb model at the 17p and the MYCN locus. We find five tumors of deceased patients to harbor extreme CN ratios due to LOH of almost the entire chromosome arm 17p (Figure 36a). Two tumors showed focal LOH at 17pter, one of which belonged to a deceased patient. Additionally, we find 9 samples with weak and strong CN imbalance due to imbalanced gains or ploidy neutral gains on 17p. We find 14 of 22 patients (64%) who

deceased from the disease to have imbalanced CN states of almost the entire chromosome arm 17p compared to 11 of 68 patients (16%) with other survival status. Thus, both LOH and gains contribute to a relatively greater proportion of samples with high CN ratio on chromosome 17p in the group of deceased patients. High CN ratios at the MYCN locus were mainly driven by focal amplifications, but we also observed two samples with LOH, one of which additionally harbored a MYCN amplification. We detected 10 out of 22 deceased patients (45%) and 5 out of 68 (7%) with other survival status to have a focal amplification at the MYCN locus. Figure 36 shows CN ratio, logR and states of CN segments at the significant 17p bins and the significant MYCN bin respectively.

We analyzed survival of patients based on the CN ratio of 17p by a Cox proportional hazard model, incorporating all covariates used in the discovery model and found both MYCN amplification and 17p copy-number ratio to be significantly associated with disease-specific survival (Figure 37a). Survival curves of donors affected by 17p imbalance were estimated using the Kaplan-Meier method. To define discrete categories, that are required by this method, we divided samples into two groups based on their copy-number ratio at 17p: Samples with copy-number ratio greater or above 0.5 were assigned the *17p imbalance state* and those with copy-number ratio 0.5 were assigned the *17p balance state*. The time-dependent survival function predicted by 17p imbalance was then visualized in a Kaplan-Meier plot (Figure 37b). Survival probability was significantly reduced (P = 0.002) for 17p imbalance compared to the 17p balance state.

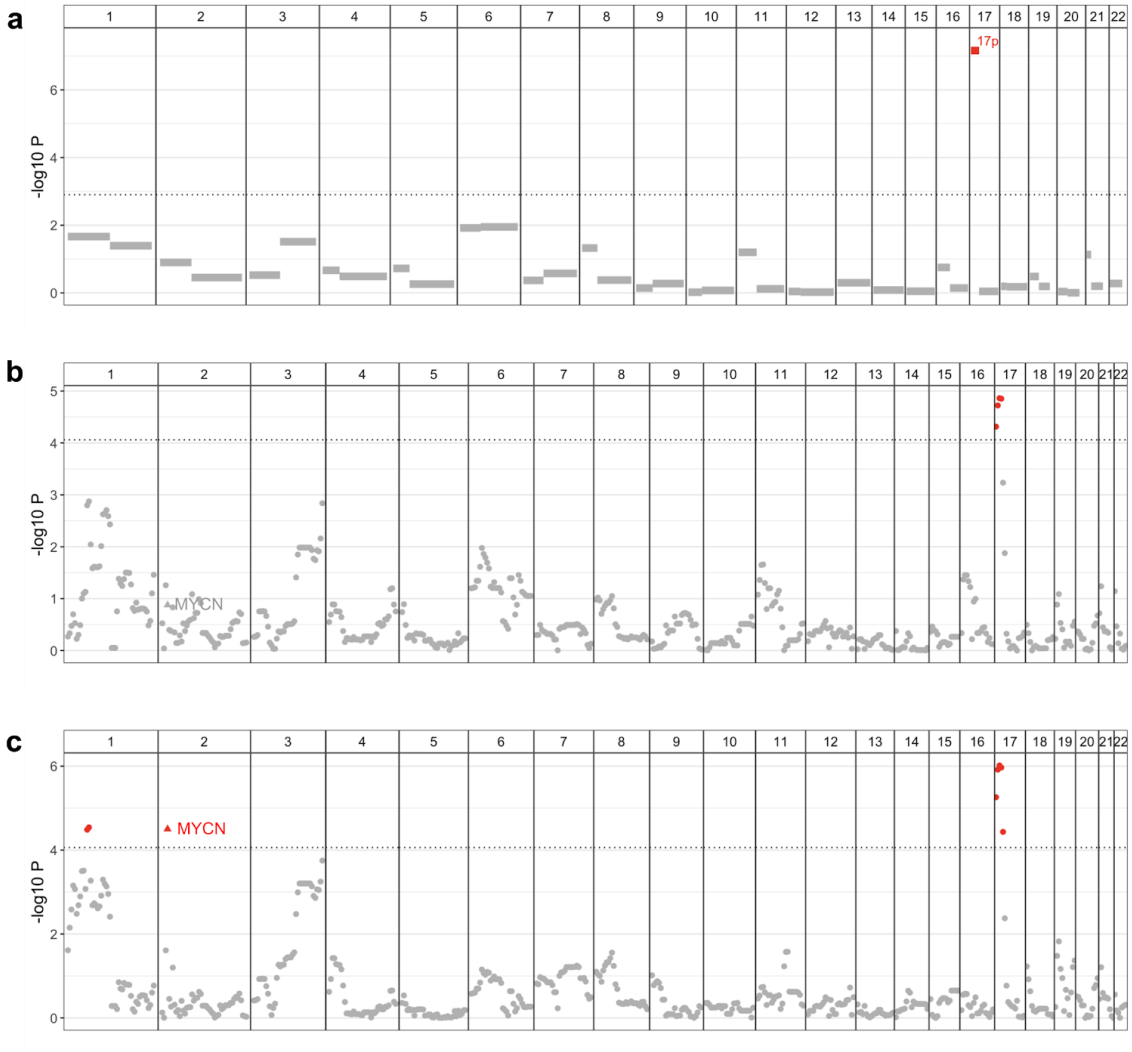


Figure 35: Genome-wide association results of copy-number ratio and survival status “deceased from disease”. Observations significant after adjusting p-value for multiple testing in red, others in grey. Significance threshold (FWER 0.05) demarcated by grey dotted line. **a**, Association on the level of chromosome arms. **b-c**, Association on the level of 5 Mb bins controlling for MYCN-amplification status (**b**) and without controlling for MYCN-amplification status (**c**). Triangles indicate overlap between the 5 Mb genomic bin and MYCN.

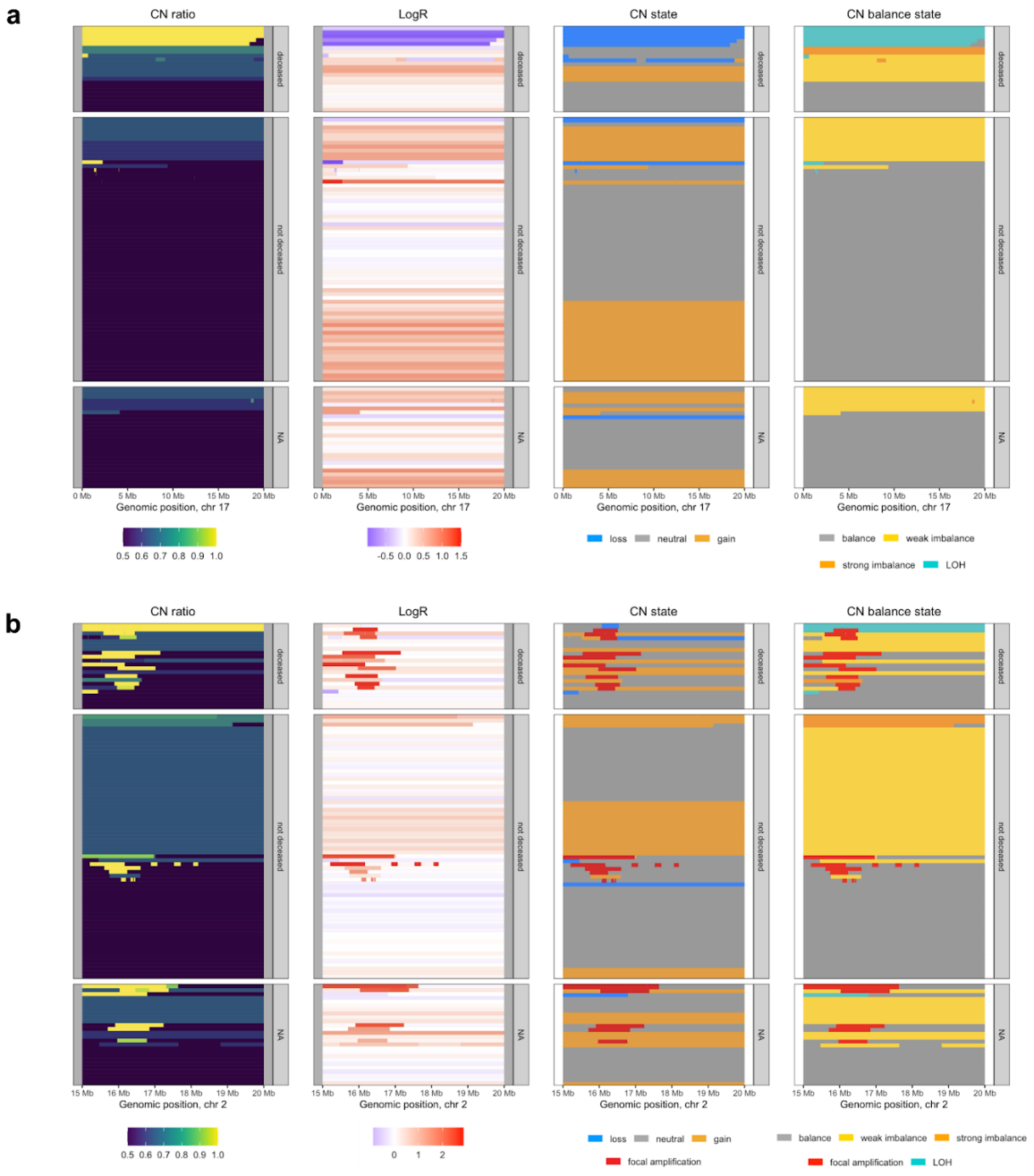


Figure 36: Copy-number observations in genomic regions associated with survival status “deceased from disease”. Observations for (**a**) chromosome arm 17p in genomic interval 0-20 Mb and (**b**) in a single genomic bin overlapping the MYCN gene on chromosome 2p at genomic interval 15-20 Mb.

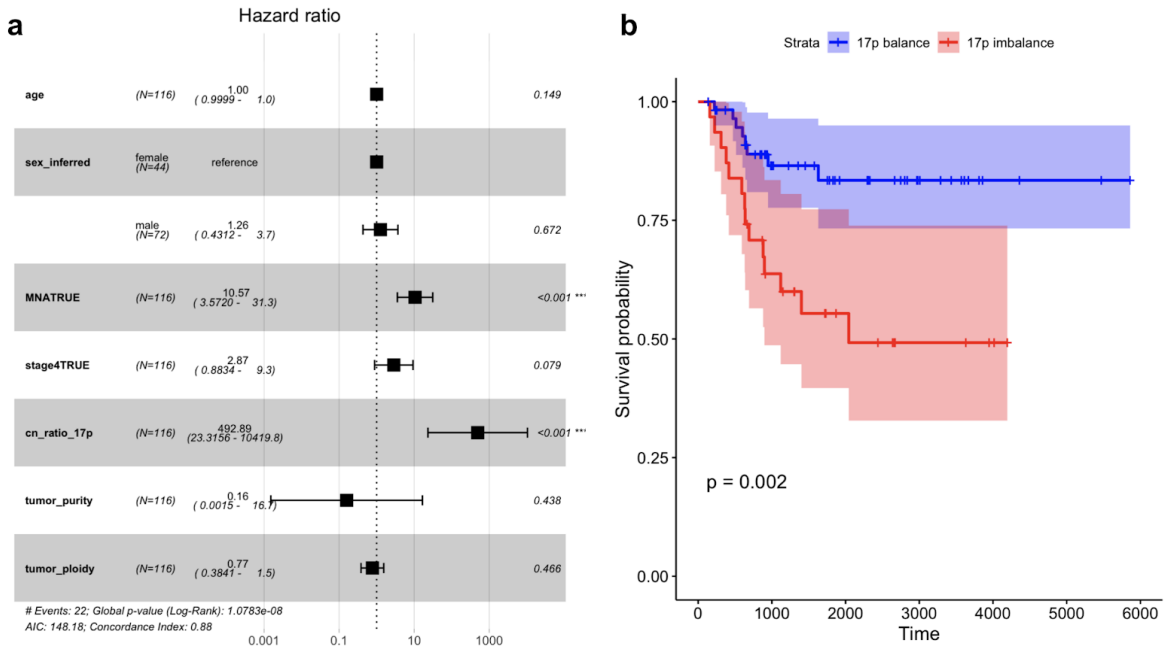


Figure 37: Hazard and survival analysis of chromosome arm 17p copy-number imbalance. **a**, Hazard ratios and p-values determined by Cox proportional hazard model for survival by copy-number ratio of 17p and additional model covariates. Hinges mark 95% confidence interval. **b**, Kaplan-Meier estimate for survival curve by copy-number imbalance on chromosome 17p. Censored data indicated by vertical marks. Colored ribbons correspond to 95% confidence intervals.

We aimed to identify gene expression consequences of 17p CN alterations that could underlie the association of 17p CN ratio and disease-specific survival. To this end we compared the CN dosage effect (Section 3.2.5) with differential expression analysis results between tumors of patients that deceased from the disease and those of other survival status (Section 3.1.3). We find 158 CN dosage effect genes and 90 differentially expressed genes on 17p, of which 60 showed both differential expression and CN dosage effect (Figure 38b). By comparing the Log2 fold-change to the expression variance explained by copy-number effects per gene, we find that almost all CN effect genes show negative Log2 fold-change or significant down-regulation in deceased patients (Figure 38a). We found ULK2, ANKFY1, MAP2K4, ZNF624, NCOR1, ALDH3A2, SMG6, C17orf85, TIMM22, ZNF287, VPS53, PAFAH1B1, SRR and LRRC48 to be down-regulated in deceased patients and to show strong copy-number dosage effect (>30% expression variance explained by LogR). PIRT showed the strongest downregulation across all genes (Log2 fold-change = -1.78) and 29% of its expression variance was explained by copy-number. Myosin genes MYH4, MYH1, MYH8, MYH2, MYH12, which are all located in a gene cluster on 17p13 and the gene SHISA6 were substantially upregulated (Log2 fold-change > 1.5). However, all but

MYH13 lacked detectable CN effects on expression. The hallmark tumor suppressor gene TP53 is located on 17p and we examined its CN dosage and differential expression. We did not detect a significant CN dosage effect on TP53 and we found weak upregulation of the gene in diseased patients (P = 0.001, Log2 fold-change = 0.55, DEseq2 test), indicating that expression of TP53 is not or only weakly affected by CN alterations and it is not down-regulated in tumors of deceased patients. We did not detect targeted somatic alterations in TP53 in 11q LOH samples. To determine gene interactions and processes associated with differentially expressed CN dosage effect genes we analyzed protein interactions and functional enrichments using the STRING database[15]. We found 18 GO terms of biological processes enriched (FDR < 0.05, STRING), several of which are associated with neurological processes (Supplementary table 14). Figure 38c shows protein network interactions of differentially expressed CN dosage genes on 17p and highlights those genes, that are involved in the processes “Regulation of neuron projection development” (GO:0010975), Chemical synapse transmission (GO:0007268) and “Regulation of neurotransmitter receptor activity” (GO:0099601).

[15] STRING DB, accessed 2021-01-06

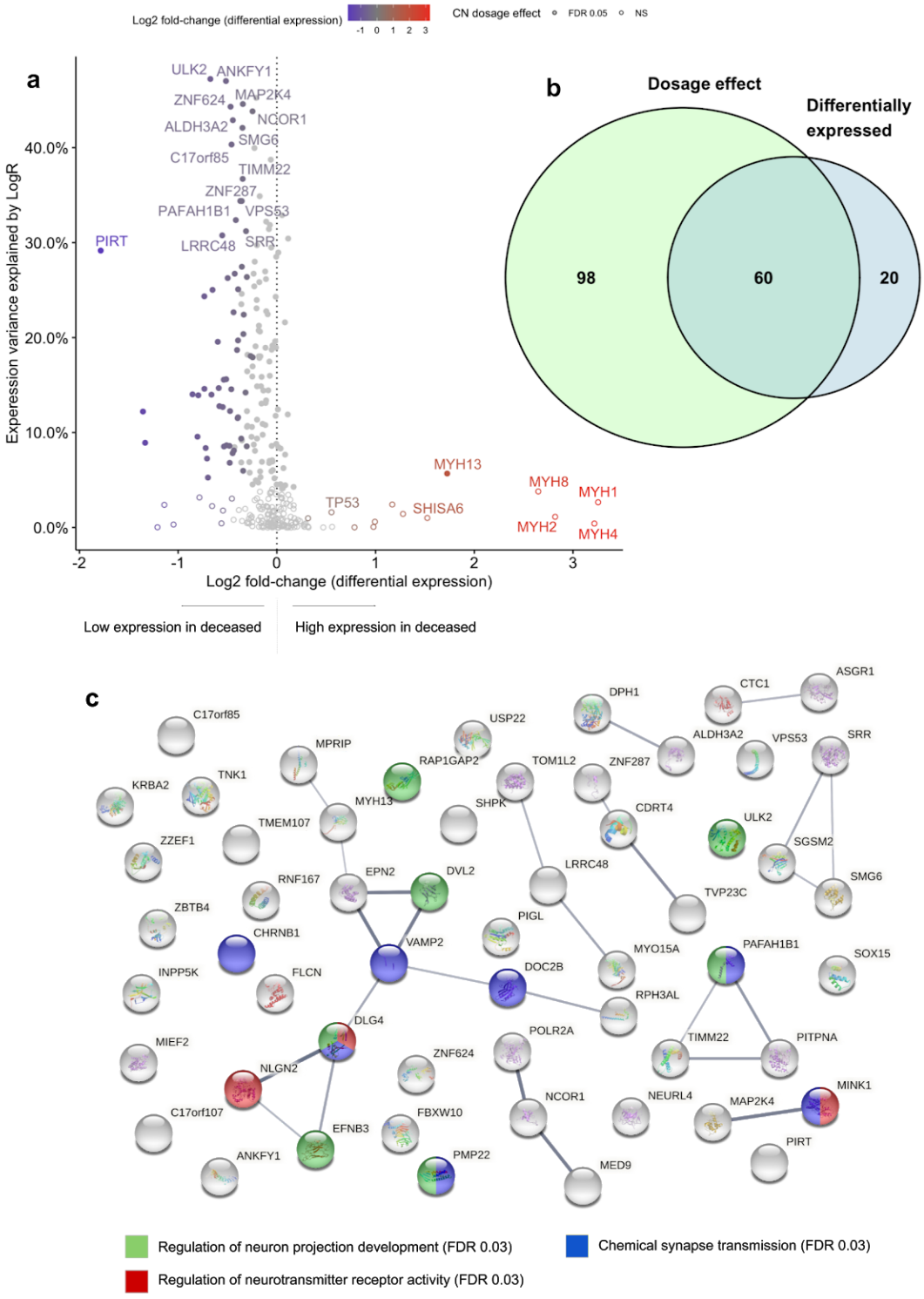


Figure 38: Differentially expression of disease-specific survival and copy-number dosage effect for genes on 17p. **a**, Expression variance explained by LogR and differential expression Log2 fold-change. Differentially expressed genes in color scale, others in light grey. CN dosage effect genes as filled circles, others as empty circles. Genes of Log2 fold-change >1.5 or those >30% expression variance explained by LogR labeled by name. TP53 is additionally labeled. **b**, Overlap of number of differentially expressed and dosage effect genes on 17p. **c**, Protein interaction network of differentially expressed CN dosage effect genes, colored by selected GO-terms indicated below.

3.3 Discussion

We here presented a systematic analysis of effects of genetic variation on gene expression across 116 neuroblastoma tumors. To this end we developed a comprehensive bioinformatics pipeline that detects germline and somatic variation based on WGS of tumor and normal tissue as well as quantification of gene expression phenotypes from RNA-seq. Its results include germline SNP calls, allele-specific copy-number calls and quantification of total and allele-specific gene expression. We established an extensive panel of coding and non-coding SNPs and identified genome-wide frequencies of somatic LOH, losses, gains and amplifications. Integrating these results with somatic SV and SNVs showed how genetic variation contributes to local quantitative expression phenotypes and how specific regulatory effects contribute to disease mechanisms.

Expression imbalance in imprinted genes

We investigated AEI frequency and strength of expression imbalance by mean ASE ratio per gene and identified a group of genes for which almost all informative samples showed AEI and harbored strong expression imbalances (Figure 19). These frequent AEI and strong expression imbalance genes were significantly enriched in genes reported to be imprinted (Morison, Ramsay, and Spencer 2005), meaning that one of the two alleles is inactivated by methylation while the other is expressed in a parent-of-origin-dependent manner (Section 2.2.1). In genes with highly recurrent AEI (> 90%) we identified very strong ASE ratios (> 0.9) in PEG10, PEG3 and IGF2, indicative of mono-allelic expression. Assuming that mono-allelic expression is caused by parent-of-origin imprinting at these gene loci, our results confirm the preservation of IGF2 imprinting in neuroblastoma tumors, in line with previous reports (Wada et al. 1995). Differential regulation of imprinted genes between neuroblastoma tumors through e.g. variable methylation levels in ICRs could play a role in cellular phenotypes and disease-associated mechanisms. Such genes would likely show lower levels of mean ASE ratios across tumors. Highly recurrent AEI (> 90%) with lower mean ASE ratios (< 0.9) were found for PLAGL1, DLK1, RTL1 and L3MBTL1, suggesting imperfect imprinting and bi-allelic expression of these genes in some of the neuroblastoma tumors investigated.

PLAG1 Like Zinc Finger 1 (PLAGL1) is a transcriptional repressor that interacts with p53 and controls cell cycle and progression through multiple converging pathways (Vega-Benedetti et

al. 2017), suggesting that imprinting-mediated transcriptional suppression of this gene could have tumor promoting effects. The Lethal(3)Malignant Brain Tumor-Like Protein 1 (L3MBTL1) is a gene from the polycomb group, which binds methylated histones and represses transcription by chromatin compaction (Trojer et al. 2007; Min et al. 2007), indicating that regulation of this gene by gain or loss of imprinting could impact expression of a variety of downstream targets due its chromatin remodelling activity. Additionally, L3MBTL1 directly binds to non-histone transcriptional regulators p53, TEL and RB and may cooperate in the repression of their downstream targets (West and Gozani 2011), so that regulation of L3MBTL1 might have direct implication in tumor suppression in cancer-associated pathways.

The RTL1 and DLK1 genes are located at approximately 145 kb distance in the imprinted DLK1-DIO3 cluster on 14q32 from which the three genes DLK1, RTL1 and DIO3 are expressed from the paternally inherited chromosome. The attenuated ASE ratios (relative to strong imprinting at the IGF2 locus) may result from a broader variable imprinting pattern at their common locus. The Non-Canonical Notch Ligand 1 (DLK1) gene encodes for a transmembrane growth-factor repeat-containing protein. DLK1 was first identified in a study of differentially expressed genes in lung carcinoma and neuroendocrine tumor cell lines (including neuroblastoma cell line SK-N-SH) (Laborda et al. 1993) and is associated with tumor formation and progression in glioblastoma (Yin et al. 2006; Grassi et al. 2020). DLK1 expression in neuroblastoma cell lines was associated with neuroendocrine lineage differentiation (Van Limpt et al. 2003), suggesting that imprinting heterogeneity in tumors could reflect a differentiation stage in premalignant cells. The retrotransposon Gag Like 1 (RTL1) gene is one of approximately 50 transposon-derived genes that were “domesticated” in the human genome (Riordan and Dupuy 2013). It is involved in placental/neonatal development (Sekita et al. 2008) but recent studies in mice show that it is also widely expressed in the nervous system (Kitazawa et al. 2021). A mutagenesis screen identified alterations that induce strong upregulation of RTL1 to confer a selective growth advantage in hepatocarcinoma cells (Riordan et al. 2013). Riordran and colleagues also showed that high RTL1 expression can promote hepatocarcinogenesis in vivo and that it is overexpressed in human hepatocarcinoma cells. In melanoma, RTL1 was found to promote cell proliferation by regulating Wnt/β-Catenin signalling (G. Fan et al. 2017). Recently, RTL1 was identified as one of 16 genes informative for survival time in high-risk neuroblastomas, with stronger RTL1 expression associated with shorter survival (Giwa et al. 2020). These results suggest

that RTL1 is a potent oncogene, which is involved in Wnt signaling, regulated through imprinting mechanisms and upregulated in some cancer entities.

In our study we have analysed genes whose ASE is associated with higher or lower levels of total gene expression. Genes for which we identified a significant correlation between these quantitative traits were termed “allelic regulated genes” (AR genes) (Section 3.1.10). We identified 467 AR genes of which 122 were also differentially expressed between patients who deceased from the disease and those who did not. Besides strongly copy-number regulated genes, like MYCN amplicon genes, other amplified genes and those in recurrent CN losses we found three copy-number independent differentially expressed AR genes including the aforementioned RTL1 (Figure 33).

Upregulation of RTL1 was associated with bi-allelic expression independent of the underlying copy-number ratio and low ASE ratios were enriched in tumors of deceased patients (Figure 34a,b). Together with previous findings of RTL1 upregulation in hepatocarcinoma and melanoma (above) our results strongly suggest that upregulation of RTL1 by loss of imprinting is associated with low survival in neuroblastoma. As we did not investigate methylation in the primary tumors directly, our results provide strong but not conclusive evidence for this phenomenon. We did not find somatic alteration that could explain deregulation of RTL1 expression. Thus, varying expression levels of RTL1 may be a result of cell lineage differentiation as reported for the DLK1 gene in the same imprinting region (see above) and loss of imprinting could be the underlying regulatory mechanism, similar as reported for IGF2 in Wilms’ tumors (Hubertus et al. 2011). Antisense microRNAs targeting RTL1 are expressed from the maternal allele and are candidate regulators of RTL1 in trans (Davis et al. 2005; Mainieri and Haig 2019). If these maternal microRNAs repress RTL1, then a switch from anti-sense to sense-transcription on the maternal allele could result in a substantial increase of RTL1 mRNA. Thus, already a subtle decrease in expression imbalance (smaller ASE ratio) could explain strong upregulation of RTL1 (Figure 33b).

As shown by Fan et al. in melanoma, upregulation of RTL1 could affect proliferation by Wnt/β-Catenin signalling in neuroblastoma, but further studies are required to identify the exact mechanism of RTL1 regulation and its functional implications. We suggest that future investigations should determine allele-specific imprinting at RTL1 and the broader DLK1-DIO1 locus in neuroblastoma and investigate its association with survival.

Furthermore, the role of maternally expressed antisense microRNAs may shed light on RTL1 regulation in these tumors.

Genomic and expression imbalance in MNA and non-MNA tumors

We determined the number of genes affected by copy-number imbalances and AEI per sample and found that these measures strongly correlate (Figure 20). Generally, samples had higher numbers of copy-number imbalance genes than genes affected by AEI. If the copy-number imbalance underlie AEI one could expect this number to be equally high. Additionally, one would expect to detect AEI in copy-number balance regions due to cis-regulatory effects of functional variation and imprinting. However, not all samples are informative for ASE in a given gene, which decreases the genes for which we can detect AEI. Additionally, the sensitivity to detect AEI is limited in cases of low allelic RNA counts and due to the multiple testing burden, as our detection method is based on a binomial test controlling for a FDR. Thus generally, we cannot exclude expression imbalances if the AEI test does not detect it and have to assume that the number of genes that are affected by expression imbalances is likely underestimated by the number of AEI genes that we identified.

Most MNA samples had fewer CN alterations as determined by our chromosome-arm level LogR analysis. Here, a large group of MNA samples characterized by 1p loss and 17q gain but otherwise very few chromosome-arm alterations formed a cluster in the high risk group (Figure 13). We therefore wanted to investigate if MNA samples in general harbor fewer CN imbalances and AEI genes. We found both the number AEI genes and CN imbalance genes between MNA and non-MNA samples so be significantly different (Figure 20b,c). These results confirm that general MNA tumors are genomically more stable and indicate that this genomic stability also leads to fewer genome-wide expression imbalances. It is very likely that amplification of the MYCN oncogene is such a potent cancer driving event that affected tumors do not require many additional alterations (Q.-R. Chen et al. 2004). Inversely, this would indicate that the deregulation of the non-MNA tumor transcriptome is mainly due to segmental and chromosomal CN alterations. However, we do observe 13 MNA tumors that cluster outside of the genomically more stable MNA group. And these “unstable” MNA tumors harbor additional alterations, such as 11q loss, 3p loss or loss of chromosome 9. Notably, only two donors from this group displayed progression free survival and the majority (N=8) donors deceased from the disease (survival status unavailable for 3 of 13 donors), which indicates that MNA tumors with CN alterations in addition to the more prevalent 1p

loss and 17q gain are at particularly high risk. Accordingly, a dramatic decline in survival rate was previously shown for MNA tumors with 11q loss (Spitz et al. 2006).

Local genetic effects on gene expression

To quantify the effects of different local genetic variants on expression of proximal genes we analyzed the contribution of observed genetic variation to the variance of total gene expression and ASE by linear regression and ANOVA (Section 3.2.3). Our results provide estimates for the relative genome-wide contributions of somatic CN, SVs and SNVs as well as germline genetic variants (Figure 21). We found a similar order of relative contributions to total expression and ASE. Highest relative contribution was attributed to somatic CN, followed by germline genetic effects, somatic SVs and SNVs, similarly as observed in a larger pan-cancer study (PCAWG Transcriptome Core Group et al. 2020). Our results show that the effect of CN alterations have profound consequences on (allelic) RNA levels and we predicted this effect to globally dominate germline regulatory effects, which we identify as the second strongest contributor. We observed a higher relative contribution of CN to ASE compared to total gene expression, which indicates that CN effects are strong local regulators of gene expression and that ASE is less affected by trans-regulation. Conversely, trans effects could lead to the high rate of unexplained variance we observed in the model of total gene expression, while unexplained variance in the ASE model could reflect unknown cis-regulatory effects. In addition both quantitative traits will also be affected by biases and measurement noise, that are also captured by the fraction of unexplained variance. We included MNA status as a covariate in both models and found that it explains relatively more total expression than ASE. Because MYCN acts as a TF in trans, this provides additional evidence that trans-regulation is more reflected by total expression than by ASE. The lower but still existing effect of MNA on ASE may be e.g. due to allele-specific binding of MNA at CREs harboring functional heterozygous variants.

Copy-number dosage effects of amplifications

To describe the effect of copy-number alterations on gene expression we classified CN segments based on imbalance and amplification status and compared associated patterns of gene expression (Section 3.2.4). We classified CN segments into the CN states *balance, weak imbalance, strong imbalance, LOH* and *focal amplification* and compared both the frequency of genes harboring AEI and the distribution of ASE ratios within these copy-number states (Figure 23). Our comparison shows that AEI frequency and average ASE ratio increase with the strength of CN imbalance, underlining results from our ASE

variance analysis that showed a pronounced effect of CN ratio on ASE (see above) and additionally that the effect is proportional to allele abundance. LOH and amplification state harbored the highest frequency of AEI observations and the strongest ASE ratios. However these two CN states have opposing effects on total RNA level when comparing the samples' expression percentiles within genes (Figure 25c). Here, expectedly, we found that samples with genes overlapping focal amplifications show the strongest expression levels of the respective gene across the cohort and genes in LOH the lowest expression, because the majority of LOH is found in CN losses. We identified amplifications of established cancer census genes ALK, BCL7A, BRCA1, CCND1, CDH1, CDK4, CLIP1, ETV4, LRIG3, MDM2, MYCN, NCOR2, PRDM1, PTPRB, RFWD3, SETD1B, ZCCHC8 and ZFHX3, showing that genetic regulation of these genes by strong CN dosage increases likely confer important properties in proliferative signaling (e.g. ALK), replicative immortality (e.g. CDK4), genome stability (e.g. CCND1) as well as invasion and metastasis (e.g. BRCA1) (Tate et al. 2019). The small CN segment size of amplifications, which are often termed "focal" because of that reason, makes it relatively easy to pinpoint the relevant target genes of these alterations. Still co-amplifications of multiple genes on the same CN segment occur regularly, as seen e.g. by co-amplifications of DDX1 and NBAS along with MYCN (Noguchi et al. 1996; Wimmer et al. 1999). Moreover, we showed that CN alterations deregulate gene expression genome-wide and these include large-sized chromosomal and segmental CN changes. Due to the large size of these alterations they encompass many more genes than focal amplifications and it is therefore more difficult to pinpoint specific gene targets.

Pathway enrichment in copy-number dosage effects

The question remains, which genes and disease mechanisms are targeted by large SCNAs. We speculated that expression levels of genes on the same copy-number segment would be affected differently by copy-number alterations dependent on the gene's sensitivity to CN dosage. This is because trans-regulatory effects might superimpose local genetic effects. Dosage-compensation, which has mainly been studied in the context of sex chromosomes (Ferrari et al. 2014), could rescue expression levels of genes affected by SCNAs in cancer. For example, a simple negative feedback loop by which a gene product stabilizes its own expression in a trans (Pastinen 2010, Figure 1) could explain reduced CN dosage sensitivity in a subset of genes. We could assume that genes which are rescued from SCNA effects are less likely to be targets of these alterations. Conversely, genes that are (strongly) affected by SCNAs could play critical roles in disease mechanisms. To investigate which genes and pathways are influenced by SCNAs, we determined the CN dosage sensitivity of

individual genes by the proportion of expression variance explained by LogR and determined pathway enrichment in CN dosage effects (Section 3.2.5). Significant effects of somatic CN on total gene expression was detected in approximately half of the genes with effects of up to 71% of expression explained by somatic CN and amplifications showed particularly strong dosage effects (Figure 26). However, we also found marked dosage effects of 50% and more variance explained in broader regions of losses and gains, indicating that somatic CN is a strong genetic regulator in a subset of genes affected by losses and gains, which introduce only moderate CN dosage differences compared to amplifications. Growing evidence suggests that larger gains and losses are specifically selected in cancer genomes because of their potential to regulate genes that are sensitive to somatic CN dosage alterations (Solimini et al. 2012; Greenman 2012; Davoli et al. 2013; Fehrmann et al. 2015; Cai et al. 2016; Sack et al. 2018; Shao et al. 2019). And many more genes than anticipated were found to regulate proliferation and the effect was often cell-type specific (Sack et al. 2018), a fact that could explain why patterns of SCNAs are often characteristic for specific cancer types. A continuum model for tumor suppression suggests that the effect of tumor suppressor genes can be mediated by gradual dosage as opposed to complete inactivation (two hit hypothesis) and that different expression levels may lead to varying phenotypic outcomes dependent on the tissue (Berger, Knudson, and Pandolfi 2011). Together these studies indicate that global patterns of SCNA are important genetic regulators subject to selection towards gains with oncogene- and losses of tumor suppressor-like properties. We investigated biological mechanisms targeted by dosage effects of SCNAs. Here, our analysis revealed enrichment of CN dosage effects in several pathways including “Cell cycle”, “DNA Repair”, “Regulation of TP53 Activity”, “MAPK6/MAPK4 signaling”, “PTEN Regulation”, “Regulation of RUNX3 expression and activity”, “NTF3 activates NTRK3 signaling” and “Regulation of MECP2 expression and activity”. Thus, our work provides evidence for biological mechanisms targeted by somatic CN alterations in neuroblastoma tumors. This form of genetic deregulation converges on important cancer-hallmark associated pathways, such as cell cycle, genome stability and repair (TP53), tumor suppression (TP53, PTEN, RUNX3) as well as pathways associated with neuronal cell differentiation (NTRK3) and epigenetic regulation (MECP2). Therefore, our findings may provide cornerstones in investigations of neuroblastoma biology, including those focusing on malignant transformation in the developing nervous system by cell differentiation defects and epigenetic reprogramming.

Copy-number gains of activated TERT

We seeked to investigate the effect of SCNAs on telomere maintenance mechanisms. To this end we examined the relation between MNA, rearrangements and SCNAs at the TERT locus (Section 3.2.7). A comparison of these observations recapitulated previous results that associated both MNA as well as TERTr with increased expression of TERT (Mac, D'Cunha, and Farnham 2000; Peifer et al. 2015; Valentijn et al. 2015). Furthermore we found SCNAs to increase TERT expression specifically in tumors which showed MNA or TERTr, and no SCNAs effect on TERT expression was found in tumors that lacked TERT activation (Figure 31). Our results show that SCNAs cooperate at this locus to increase TERT expression in tumors that acquired this form of telomere maintenance. Interestingly the results imply that copy-number gains selectively target TERTr alleles, assuming that only one of the alleles is rearranged. However, we here did not provide further evidence, as this would require ASE at the TERT locus. Unfortunately, most samples in our cohort were uninformative for ASE in TERT, likely because of a lack of expressed exonic hetSNPs. We can speculate that copy-number gains at this locus might be selected for in TERT activated samples if increased expression provides a more effective escape from cellular senescence (Section 2.4) than a "weak" TERT activation alone. Furthermore our findings provide evidence for an interplay between SCNAs with both trans-regulatory factors (MNA) and cis-regulatory alterations (TERTr). One could speculate that such dependencies may shape SCNA selection in wider parts of tumor genomes. For example, embryonic tumors that originate from different developmental stages of progenitor cells likely have diverging epigenomes and regulatory programs. If CN dosage effects are selected, then these cells will acquire a different set of SCNA alterations during tumor evolution, dependent on which genes are activated or repressed in the progenitor cells. Thus SCNA patterns will depend on the premalignant regulatory landscapes (Sack et al. 2018), which could explain a large part of SCNA heterogeneity between tumors of the same cancer and even more so across different cancer types.

Links between 11q loss, histone variants and alternative lengthening of telomeres

In order to investigate genetic determinants of ALT, the TERT-independent telomere maintenance mechanism (Section 2.4), we associated ATRX alterations and genome-wide patterns of SCNAs with telomere length (Section 3.2.6). Here, ALT was defined by an excess of telomere length in tumors vs. normal tissue. We confirmed that ALT is significantly associated with ATRX alterations. However, in the majority of ALT tumors we did not detect ATRX alterations, suggesting that in some ALT tumors the pathway is induced by other

means. Strikingly, our copy-number analysis revealed an association of ALT and loss of 11q (Figure 28), providing evidence that SCNAs are associated with ALT. Furthermore we identified differentially expressed genes between ALT and non-ALT tumors. Among highly significant upregulated genes we find the two histone variant genes H3F3B and H2AFJ. Among highly significant downregulated genes we found RAC1 (located on 7p) and CCDC90B, PPME1 and NCAM1, which are all located on 11q. Strikingly we found upregulation of H3F3B and H2AFJ to correlate with 11q loss, even though these genes are located on 17q and 12p respectively (Figure 29b-e), suggesting that these histone genes are affected by trans regulatory mechanisms linked to 11q. Comparison of ASE ratios provided further evidence for local genetic regulation of 11q genes and trans effect on the two histone genes (Figure 30a). We can speculate that upregulation of histone variant genes is caused by loss of a repressive regulatory factor, that is encoded on 11q and sensitive to the decrease in CN dosage.

Protein network analysis of ATRX and differentially expressed genes showed that H3F3B interacts with ATRX as well as H2AFJ and H3F3C, a third histone variant that we found to be upregulated in ALT tumors (Figure 30b). Histone variants can replace canonical histones in nucleosomes (Section 2.2.1). H3F3B and its paralog H3F3A encode for the same histone variant H3.3 (Frank, Doenecke, and Albig 2003). These genes are altered in several pediatric cancers by activating mutations. While H3F3B harbors driver mutations in 95% of chondroblastomas (Behjati et al. 2013), mutations in its paralog H3F3A are a hallmark of diffuse intrinsic pontine glioma and prevalent in pediatric glioblastoma multiforme (G. Wu et al. 2012; Schwartzentruber et al. 2012). H3.3 histones are deposited by the ATRX/DAXX complex at telomeres, which is predicted to stabilize chromatin to prevent replication stalling and disruption of this process was proposed to cause ALT (Section 2.4 and Clynes et al. 2013). Schwarzentruber and colleagues found that the identified gain of function mutations in H3F3A together with alterations in ATRX, DAXX or TP53 were significantly associated with ALT in paediatric glioblastomas and a recent study concluded that H3F3A mutations can trigger ALT independent of ATRX status (Minasi et al. 2021). Another mechanism described for H3.3 in eukaryotes is the displacement of canonical H3 histones in actively transcribed genes (Ahmad and Henikoff 2002; Schwartz and Ahmad 2005) and the maintenance of epigenetic memory at CREs (Ng and Gurdon 2008; P. Chen et al. 2013; Fang et al. 2018). While ATRX is required for the deposition of H3.3 at telomeres it is not essential for its deposition at transcribed genes and CREs (Goldberg et al. 2010). Thus, upregulation of H3.3 in neuroblastomas tumors harboring ATRX loss-of-function could favor

H3.3 deposition at actively transcribed genes over its deposition in telomeric chromatin. We also found 11q loss in ALT tumors lacking alterations in ATRX. And in these tumors H3.3 could could together with ATRX/DAXX still facilitate telomeric stability, if this process is not disrupted otherwise. However, it is unclear if elevated levels of H3F3B in neuroblastoma could affect ALT similarly as mutant H3F3A in gliomas.

Similar to H3F3B, the two other ALT-upregulated histone genes H2AFJ and H3F3C also encode for histone variants that can displace canonical histones and thereby alter the epigenetic state through chromatin remodeling. Histone variant H3.5 encoded by H3F3C was found to destabilize nucleosomes and accumulate at TSS in testis (Urahama et al. 2016), suggesting that its overexpression could result in a more permissive state for TF binding at promoters affected by the displacement. Histone variant H2A.J encoded by H2AFJ can displace H2A histones. It was found to accumulate in senescent human fibroblasts that are affected by DNA damage and positively regulate inflammatory pathways with potential pro-tumorigenic effect (Contrepois et al. 2017). H2AFJ was shown to be differentially expressed in cancer. An early differential expression study in melanoma found H2AFJ to be downregulated in melanoma metastasis lesions compared to nevus tissue samples (de Wit et al. 2005), while it showed copy-number induced upregulation in breast cancer (J. Yao et al. 2006). A more recent study found H2AFJ upregulation to be associated with radiation-resistant and worse survival in colorectal cancer (Xiaojie Wang et al. 2019).

None of the highly significant 11q downregulated genes (CCDC90B, PPME1 and NCAM1, RAC1) are reasonable candidates for TFs or co-factors which could repress expression of histone variant genes in trans. However, the set of all down-regulated genes on 11q might still provide suitable candidates for one or more trans acting repressors. Pathway information could be utilized to pinpoint candidate regulators by inference of protein activity levels (Alvarez et al. 2016). A regulatory effect on H3F3B and H2AFJ could then subsequently be validated by siRNA knockdown experiments. Similarly, CRISPR knockdown experiments of multiple gene candidates could be used to identify 11q regulators.

Other ATRX-interacting and differentially expressed genes in ALT tumors include the downregulated mismatch repair gene PMS2 located on chromosome arm 7p. Interestingly, low tumor DNA coverage on 7p showed nominal significance (P=0.008, ANOVA) in our SCNA–ALT association test (Supplementary table 12), indicating that ALT-associated DNA repair defects could be mediated by PMS2 expression sensitivity to 7p loss. Infact PMS2

knockdown increased the lifespan of telomere deficient mice (Siegl-Cachedenier et al. 2007). In the light of these findings our results suggest that mismatch repair by PMS2 could be involved in telomerase-independent maintenance of telomeres in neuroblastoma.

Our findings implicate 11q loss and 11q-linked upregulation of histone genes in ALT in neuroblastoma. Additionally, we found evidence for downregulation of a mismatch repair gene (PMS2) possibly linked to 7p loss in these tumors. In the context of the aforementioned studies our results suggest that upregulation of histone variants facilitates chromatin remodeling in neuroblastoma ALT tumors and that 11q loss provides a regulatory context favoring specific histone variants that are involved in telomere DNA stability, transcriptional regulation and pro-tumorigenic inflammatory pathways. Our work strongly suggests that CN dosage effects regulate ALT gene expression. Further investigations are required to better understand the role of deregulation of histone variant genes and CN dosage effects in ALT.

17p copy-number imbalance and survival

The results we presented underlined the importance of somatic CN in the regulation of disease mechanisms in neuroblastoma. Our allele-specific pipeline allows us to determine copy-number imbalances based on heterozygous SNPs from WGS at unprecedented resolution. We thus investigated the impact of CN imbalances on patient survival. To this end we analyzed disease-specific survival in relation to CN imbalance (Section 3.2.9) and confirmed that focal CN imbalances at the MYCN are associated with disease-specific mortality and are linked to imbalances on 1p. More importantly, we identified an association between 17p imbalance and disease-specific mortality (Figure 35). The association was robust when controlling for MNA, indicating that this risk-associated imbalance of 17p is independent of MYCN amplification status. We identified 5 donors with strong CN imbalance due to 17p LOH, all of which deceased from the disease, suggesting that loss of 17p minor allele underlies the association. However, we also found a substantial number of samples with weak and strong imbalances without LOH that contributed to the association (Figure 36a), suggesting that a general imbalance instead of LOH might underlie increased risk. The association test at 5 Mb resolution was still sensitive enough to detect the association despite a higher multiple testing burden. A Cox proportional hazard model revealed a significant hazard ratio for 17p imbalance, likely driven by the strong imbalances of deceased carriers of 17p LOH. Notably the estimated hazard ratio was higher than the one predicted for MNA (Figure 37), which may be due the fact that a subset of patients with MNA tumors survived, while all donors of tumors harboring 17p LOH deceased. Survival

curves obtained by a Kaplan-Meier estimator were significantly different with lower survival probabilities in cases with 17p imbalance compared to other samples (Figure 37b), confirming results from our discovery model.

We seeked to understand if copy-number dosage drives gene expression on 17p in deceased samples. Comparison of CN dosage effects to differential expression in deceased patients revealed that most 17p differentially expressed genes showed dosage effects and that these dosage effects were predominantly found in down-regulated genes (Figure 38a), suggesting that dosage-dependent downregulation of a subset of 17p genes is associated with higher mortality. Dosage effects of differentially expressed genes were enriched in neuronal development pathways, indicating that 17p losses could impede differentiation towards a neuronal cell fate in tumor cells, similar to previous reports on disruptions of neuronal genes by somatic SVs in neuroblastoma (Molenaar, Koster, et al. 2012). Strongest downregulation was found in the PIRT gene for which around 30% of expression was explained by CN dosage. Thus, PIRT is repressed by CN and another CN-independent mechanism. Interestingly PIRT was previously associated with hypermethylation in deceased patients (Olsson et al. 2016), indicating that repressive methylation patterns could explain lower levels of CN dosage effect despite stronger downregulation (Figure 38a). We did not detect a dosage effect or targeted somatic alterations in TP53, indicating that 17p loss does not substantially reduce its expression, and that it may not necessarily contribute a “second hit” in homozygous inactivation of this important tumor suppressor. However, as the sample represents a biopsy at an early time point single-copy 17p could predispose to a complete inactivation of TP53 by a later mutation on the remaining allele, which could then lead to increased genomic instability and consequently to relapse and death in these patients. Therefore single-copy 17p could still represent the “first hit” in this process. However, our finding on CN dosage-dependent down-regulation of neuronal genes suggests that the reason for the association might be more complex. A previous study showed that TP53 deletions that encompassed additional genes on 17p lead to more aggressive phenotype in lymphoma and leukaemia (Y. Liu et al. 2016). Interestingly the model of 17p13 deletion used by Liu and colleagues showed to cause accelerated lymphoma development also includes VAMP2, a dosage effect gene involved in synaptic vesicles trafficking which we identified as differentially downregulated in deceased patients (Figure 38c and Supplementary table 13).

Our results implicate 17p CN imbalance in decreased survival in neuroblastoma and show that CN dosage effects on this chromosome arm are linked to down-regulation of genes involved in neuronal development and activity. Our analysis falls short to explain how weak and strong imbalances contribute to this association, as these alterations do not necessarily induce CN losses.

Summary

In summary, we have here characterized the local regulatory impact of germline and somatic variation in neuroblastoma tumors and its association with selected disease phenotypes. Our results confirmed global differences between genomic- and expression imbalances between MNA and non-MNA tumors and a marked expression imbalance in a subset of known imprinted genes. We quantified the relative local genetic influence of germline and somatic variation and found somatic CN to dominate effects on both gene expression and particularly on ASE, providing strong evidence for its local regulatory effects. The most pronounced copy-number-associated regulatory effects identified were introduced by amplifications that lead to marked dosage-dependent upregulation from amplified alleles. We showed that copy-number alterations regulate dosage sensitive genes with effects enriched in cancer-associated pathways. Furthermore, our analysis established a mechanistic link between dosage-effects and telomere maintenance in two distinct pathways: Firstly, we showed how 5p gains cooperate with TERT activation to increase its expression. And secondly, we showed that 11q loss-dependent upregulation of histone genes is associated with ALT. We examined risk-associated allelic regulation and identified CN-independent effects for the imprinted gene RTL1, indicating that its upregulation by loss of imprinting may be linked to an unfavorable prognosis. Lastly, we showed that CN imbalance of 17p is associated with disease-specific mortality and described dosage-dependent downregulation of neuronal genes on this chromosome arm in deceased patients. Taken together, our results provide a detailed description of genetic regulation and its association with disease mechanisms and patient survival. And they underline the importance of somatic CN alterations in genetic deregulation in neuroblastoma.

4 Allelic dosage effects of extrachromosomal circular DNA

In this chapter I will investigate genome-wide patterns of ecDNAs in relation to somatic CN and gene expression. To that end I will identify haplotypes of ecDNA in 16 primary tumors by phasing of hetSNPs using statistical approaches based on SNPs and genotypes in the broader population, as well as somatic CN phasing, which uncovers haplotypes in regions of CN imbalance in tumor samples. In an integrative allele-specific analysis of Circle-seq, WGS and RNA-seq I will study ecDNA haplotypes and associated patterns of ASCN and ASE. In order to better understand how circular DNA of different sizes relates to somatic CN, ecDNA length will be associated with CN states and the overlap of CN segments will be compared with circularized genomic regions. Finally, I will present an analysis of amplification-associated ecDNAs, their allelic imbalances and effects on expression of individual genes in one of the tumors in greater detail.

Contributions to this chapter

Alignments of normal WGS, tumor WGS, tumor RNA-seq were created by the Core Unit Bioinformatics (CUBI) of the Berlin Institute of Health (Berlin, Germany) under supervision of Dr. Dieter Beule. Richard P. Koche, PhD (Memorial Sloan Kettering Cancer Center, New York, USA) and Dr. Anton Henssen (Charité – Universitätsmedizin Berlin, Germany) provided alignments of Circle-seq reads and ecDNA regions per sample. Parts of this chapter have been published in Koche et al. 2020.

4.1 Methods

4.1.1 Sample preparation and sequencing

Preparation sequencing and alignment of Circle-seq samples was performed in a subset of tumors from NB2004 donors as described earlier (Koche et al. 2020). In brief, circular DNA isolation and purification was performed on primary tumor samples similarly to the report of circular DNA in yeast by Circle-seq (Henrik D. Møller et al. 2015). Resulting Circle-seq libraries were sequenced on MiSeq instruments with 2 × 150 bp paired-end reads, HiSeq 4000 instruments with 2 × 125bp paired-end reads, or NextSeq instruments with 2 × 150 bp

paired-end reads (Illumina, San Diego, USA). Reads were aligned to the human reference assembly GRCh37 (hg19) with BWA-MEM 0.7.15 (H. Li and Durbin 2009) and optical duplicates were removed. For 20 subjects matched Circle-seq, WGS and RNA-seq from tumor and WGS from the blood was available. These samples were subject to the analysis described in this chapter. Supplementary table 1 lists donor tumors, from which Circle-seq data was obtained.

4.1.2 Identification of circularized genomic regions

Circularized genomic regions were identified by our collaborators as described in (Koche et al. 2020). In brief, regional enrichment of Circle-seq reads were identified by peak calling. Boundaries of resulting peaks were examined for overlapping split reads or read pairs with outward facing orientation (circle-supporting reads). A minimum threshold for the number of circle-supporting reads was determined from an empirical background distribution of WGS circle-supporting reads in regions that did not overlap the Circle-seq peaks and a one-sided empirical $P < 0.01$. Peaks with a number of circle-supporting reads overlapping its boundaries above this threshold were classified as circularized genomic regions.

4.1.3 Allele-specific expression analysis of circles

Allele specific expression of heterozygous SNPs was determined as described in section 3.1.5. RNA-seq B-allele frequencies at hetSNPs were calculated in the same manner as described for WGS in section 3.1.6. We phased SNPs based on a combination of statistical phasing (Section 3.1.4) in balanced CN regions and copy-number phasing in CN imbalanced regions (Jamal-Hanjani et al. 2017) . In copy-number phasing SNP alleles in imbalance CN region with the same direction of deviation from 0.5 in tumor WGS BAF are assigned to the same haplotype. We determined haplotype ASE counts per circle by summing over counts of phased alleles. Allelic expression preference of mono-allelic circles in copy-number balanced regions was determined by a statistical test on the circles' ASE haplotype state. Circles in copy-number imbalanced regions (Section 3.1.6) were removed. From the remaining circles only mono-allelic circles were retained, which were defined as circles with Circle-seq maximum haplotype frequency > 0.9 (Section 4.1.4). Circles with identical RNA counts for both haplotypes were removed and remaining circles were annotated with two different states dependent on whether the majority of ASE counts came from the circularised haplotype or not. Finally a binomial test on the circles was conducted, parameterized for equal probability to draw a circle from one of the two states under the null hypothesis.

4.1.4 Assignment of CN states to circles

Copy-number segments from allele-specific copy-number analysis of tumor and normal WGS samples were calculated as described in section 3.1.6. Copy-number states of segments were defined as follows: *Balance*: total copy-number ≧ 0 and majorCN = minorCN, where majorCN and minorCN are the copy-numbers of major- and minor allele respectively; *Weak imbalance*: total copy-number ≧ 0 and copy-number ratio = majorCN / (majorCN + minorCN) ≦ ⅔; *Strong imbalance:* as weak imbalance, but copy-number ratio > ⅔; *LOH*: minorCN = 0 and majorCN > 0; *Focal amplification*: copy-number segment smaller than 3 Mb, logr_seg > 0.8 and median(logr_seg) - median(logr_chr) > 0.8, where logr_seg and logr_chr are coverage log ratios of SNPs on the segment and its chromosome of origin respectively. The segment of largest overlap with a circularized genomic region was identified and its copy-number state was assigned to the circle.

We phased heterozygous SNPs in regions of imbalanced copy-number based on their tumor BAF (Section 3.1.4 and (Jamal-Hanjani et al. 2017)). The phase of SNPs overlapping imbalanced copy-number segments (majorCN > minorCN) was defined such that haplotype 1 was assigned to the minor allele and haplotype 2 to the major allele according to the tumor WGS BAF at that SNP. Only heterozygous SNPs with a Circle-seq coverage of 10 or more reads and circles with at least one SNP fulfilling this requirement were included in the allele-specific analysis. Circle-seq haplotype counts were defined as the sum over allelic-depths of the same haplotype for SNPs overlapping the circle. The haplotype frequency was calculated as hc2 / (hc1 + hc2), where hc1 and hc2 are haplotype counts of haplotype 1 and 2 respectively. The maximum haplotype frequency per circle was calculated as max(hc1, hc2) / (hc1 + hc2).

4.1.5 Circle length analysis

To identify length preferences for circles depending on the copy-number state of the underlying genomic segment we derived a zero-sum score, following common enrichment test strategies such as Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005). For a given CN category (*balance*, *weak imbalance*, *strong imbalance*, *LOH* and *focal amplification*) each circle was assigned a score of 1/*k* if the circle belonged to the category and -1/(*n*-*k*) otherwise, where *k* is the total number of circles in that category and *n* is the total number of circles. Circles were ranked by their length and cumulative scores along the list were calculated. The absolute maximum cumulative score

was tested against 10,000 random permutations of the ranked list to determine approximate enrichment p-values.

4.2 Results

4.2.1 Circular DNAs are mono-allelic

To investigate if ecDNA originates from a single allele we determined Circle-seq BAFs at heterozygous SNPs and aggregated them to haplotype frequencies in circularized regions. We then compared Circle-seq- to WGS allele frequencies obtained for the same SNPs and regions. We determined BAFs for Circle-seq and WGS at SNP positions from the 1000 Genomes project. Germline variants were phased using a combination of statistical and copy-number phasing and allelic read counts were aggregated to haplotype-level read counts for each circle (Section 4.1.3). We found remarked differences between BAF in Circle-seq and WGS: Circle-seq BAFs were almost exclusively close to zero or one and WGS BAFs were symmetrically distributed around 0.5 (Figure 39a), suggesting a dominantly bi-allelic origin of DNA sequences sampled from chromosomal DNA and a mono-allelic origin of reads in ecDNAs. We determined the circle haplotype frequency in both assays by aggregating BAFs of hetSNPs in individual circles. We then examined the distribution of haplotype frequencies in the context of Circle-seq coverage. As the frequency of the dominant haplotype increased for higher coverages in Circle-seq, we did not observe an increase in WGS, indicating that bi-allelic circles could be a result of sampling noise as they disappeared after increasing minimum Circle-seq coverage (Figure 39b). We seeked to understand how the BAFs of the two assays relate and compared their distributions across hetSNPs. SNPs with WGS BAF distributed around 0.5 showed extreme frequencies in Circle-seq and in a subset of SNPs extreme BAFs showed the same direction of skew towards extreme frequencies (Figure 39c), indicating that ecDNAs from balanced copy-numbers originate from a single haplotype and that in regions of extreme CN imbalance the more abundant haplotype dominates the origin of ecDNAs. In balanced copy-number regions we relied on statistical phasing to assign alleles to haplotypes. We investigated the direction of extreme Circle-seq BAFs in balanced CN regions in terms of the assigned haplotype. We find that statistical phasing of SNPs in circles from copy-number balanced regions separate Circle-seq allele frequencies (Figure 39d), confirming that circles originate from only one of the two alleles. These findings indicate that circularized genomic regions have a single genotype per SNP and SNP genotypes in these regions are from the same haplotype. We conclude that circular DNAs are of mono-allelic origin.

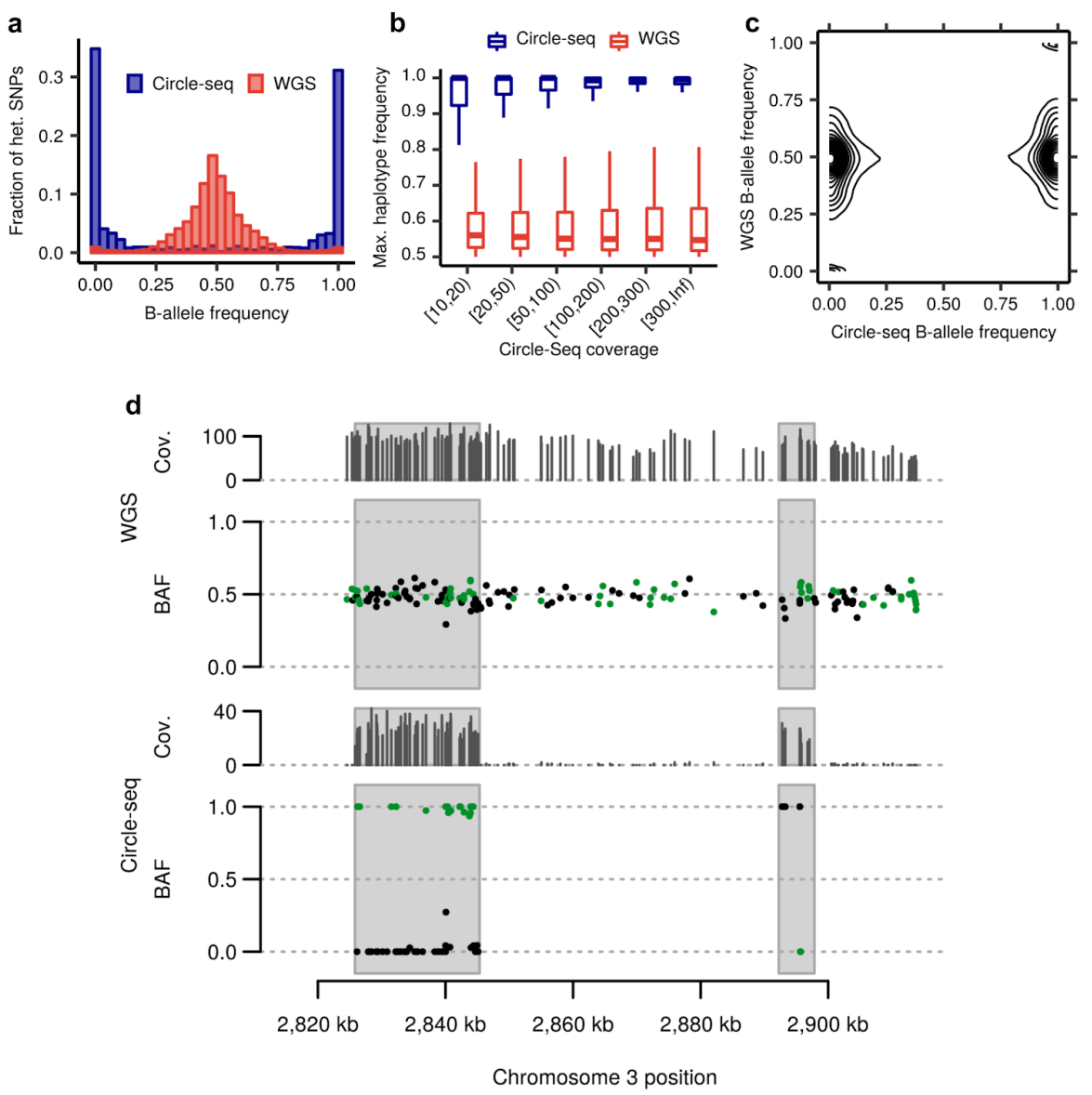


Figure 39: Circle-seq reads are mono-allelic. **a**, Comparison of B-allele frequencies at heterozygous SNPs overlapping reads in Circle-seq and WGS. **b**, Haplotype frequencies of Circle-seq and WGS in regions of circularized DNA by coverage in Circle-seq. **c**, Density of Circle-seq and WGS B-allele frequency for heterozygous SNPs in regions of circularized DNA. **d,** Example of a balanced copy-number region with coverage and B-allele frequencies determined at heterozygous SNPs (points) in WGS (top) and Circle-seq (bottom) of sample CB2008. Two distinct ecDNAs were detected (grey intervals). SNPs in which the B-allele is assigned to the same haplotype, based on statistical phasing, share the same color (green or black). Material from: Koche et al., Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma, Nature Genetics, published 2020, Springer Nature Limited'. Author reuse.

4.2.2 Somatic copy-number determines frequency of allelic origin of circular DNAs

To investigate the relationship between circle haplotype-of-origin and copy-number we assigned each circle one out of the five copy-number states *balance*, *weak imbalance*, *strong imbalance*, *LOH* and *focal amplification* (Section 4.1.4). In all imbalance states we defined the haplotype frequency per circle with respect to the major allele, so that a haplotype frequency > 0.5 indicates that the majority of Circle-seq reads are from the major allele and a frequency < 0.5 indicates that the majority of Circle-seq reads is from the minor allele. Haplotype frequencies were binned in the intervals, (0–0.1], (0.1–0.9], (0.9–0.1.0], in order to assign the origin of an ecDNA to the minor allele, both alleles (mixed) or the major allele respectively. We then compared the total amount of circles and the distributions of binned haplotype frequencies across the five copy-number states (Figure 40a and b). Across all samples investigated we found the majority of ecDNAs to in copy-number balance (48,697), followed by weak imbalance (20,702), LOH (2,028), strong imbalance (1,898) and focal amplifications (17). Mono-allelic circles in balanced regions did not show preference for any haplotype. In weakly and strongly imbalanced regions circles showed a preference for the haplotype of the major allele proportional to the degree of imbalance. As expected, we only rarely found circles from the minor haplotype in LOH, as this allele is lost in most of the tumor cells. Remaining circles of low haplotype frequencies in LOH regions could be due to subclonality of LOH events. We found a smaller proportion of circles with haplotype frequencies between 0.1 and 0.9, a range in which we would not confidently assign an allele to the ecDNA. The amount of circles with haplotype frequencies in this range was highest in strong imbalances (Figure 40b). Strong imbalances encompass regions of higher total copy-number, suggesting that in strong gains ecDNAs from both alleles with overlapping genomic coordinates exist, or more likely, that Circle-seq samples non-circularized genomic DNA in proportion to the abundance of chromosomal DNA. Our results suggest that ecDNA haplotype frequency is mainly determined by copy-number imbalances and we conclude that outside of focal amplifications, circles are not causally related to the observed gain in copy-number. In contrast, in genomic regions affected by focal amplifications, circles seemed to be exclusively derived from the amplified haplotype, as all ecDNAs in focal amplifications showed haplotype frequencies above 0.9. Thus, as expected, copy-number increases introduced by focal amplifications could still be linked to circularization.

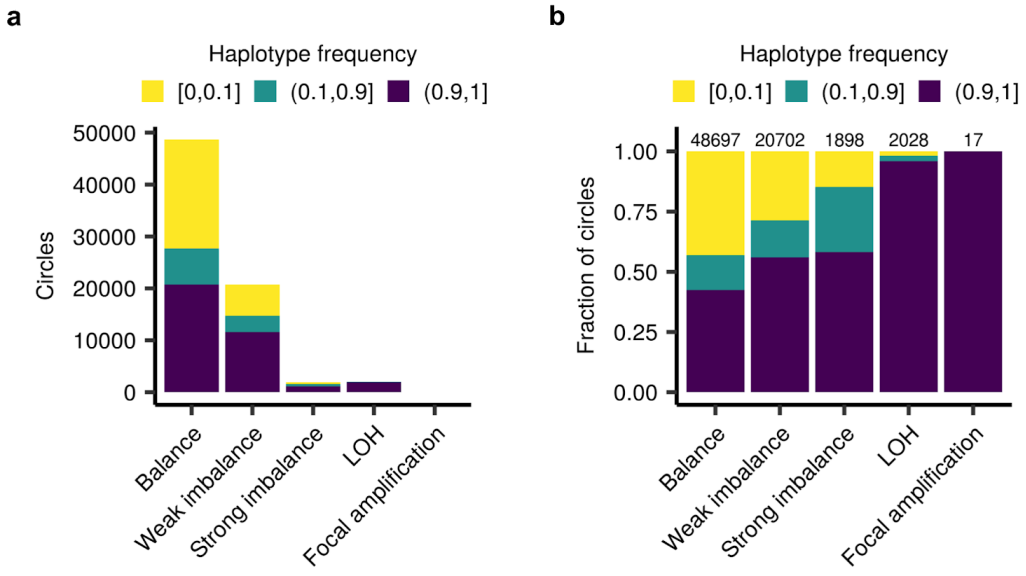


Figure 40: Circle-seq haplotype frequencies by copy-number state. Total number of circles (**a**) and fraction of circles (**b**) by binned frequency of the major allele haplotype as determined by copy-number phasing. Yellow: Circle-seq haplotype corresponds to minor allele. Violett: Circle-seq haplotype corresponds to major allele. Major/minor allele assignment in balanced copy-number state is arbitrary. Material from: Koche et al., Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma, Nature Genetics, published 2020, Springer Nature Limited'. Author reuse.

4.2.3 Large but not small ecDNAs are associated with focal amplifications

To determine the relation between circle size and copy-number we looked for enrichment of copy-number states among circles of similar lengths and compared the size and genomic coverage of copy-number segments with circular DNA overlap. We ranked ecDNA calls by their size and calculated a cumulative enrichment score for each of the five copy-number states within that ranked list (Section 4.1.5). Balances and weak imbalances showed little preference for circles of a certain length. Circles in strong imbalances tended to be longer than 10 kb, whereas circles in LOH showed a slight enrichment between 1 and 5 kb. Focal amplifications were strongly associated with the longest circles and ecDNAs longer than 200 kb exclusively originated from focal amplifications (Figure 41a). We tested the absolute cumulative score per copy-number state for significance by permutation testing and found all states to be significantly associated with circle length (balance: $P = 1 \times 10^{-4}$, other: $P < 1 \times 10^{-4}$). We argued that any causal connection between DNA circularisation and focal amplifications should lead to a strong correspondence in genomic intervals of copy-number

segments of focal amplifications and ecDNA with shared breakpoints between the circularised and amplified genomic segments. We thus determined the ecDNA overlap in CN segments and found small CN segments of extreme coverage in tumor WGS to have the highest ecDNA overlap (Figure 41b), underlining the connection between ecDNA and focal amplifications. We specifically examined boundaries of ecDNAs, read coverage and BAFs at the MYCN locus in MNA tumors and found a striking correspondence of ecDNA boundaries and genomic regions of high coverage and imbalanced BAF in tumor WGS, Circle-seq and RNA-seq. Figure 41c shows read coverage and BAF at hetSNPs for the three sequencing assays in tumor CB2013 at the MYCN locus. This tumor harbors a MYCN amplification linked to ecDNA, clearly visible by a high read coverage in tumor WGS and strong BAF imbalances in all three sequencing assays. The applied phasing in CN imbalances is based on tumor WGS BAF (Section 4.1.3) and in this figure the haplotype inferred by this method is indicated by the blue and red coloring of SNPs (Figure 41c, top). Notably, inside the amplified genomic region the direction of deviation from BAF 0.5 is consistent for SNP alleles phased to the same haplotype between tumor WGS, Circle-seq and RNA-seq (Figure 41c). Thus we conclude that in this sample exclusively the amplified MYCN allele is circularized and highly expressed. We observed that the entire amplified region overlaps ecDNAs calls (see grey boundary boxes in Figure 41c), which suggests that the entire amplified DNA sequence resides on ecDNA. Taken together our results show that large circular DNA but not smaller circles are associated with focal amplifications and they confirm that ecDNA from the MYCN locus contains a single amplified and highly expressed allele.

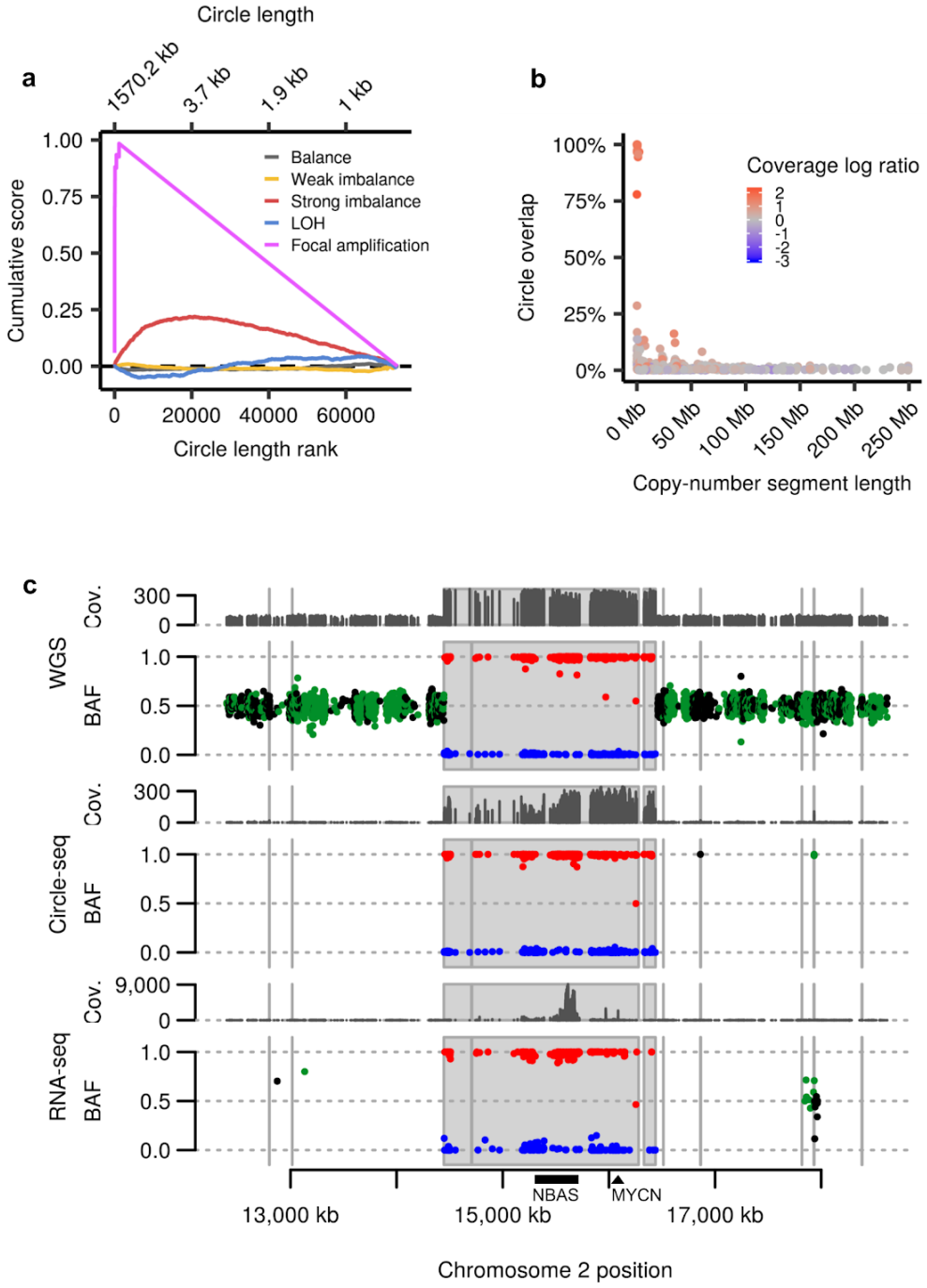


Figure 41: Focal amplifications are enriched in large ecDNAs. **a**, Enrichment of ecDNA length in copy-number states. **b**, Length, ecDNA overlap and tumor WGS coverage per copy-number segment. **c**, Comparison of coverage and B-allele frequencies (BAF) at heterozygous SNPs in WGS, Circle-seq and RNA-seq at the MYCN locus in MYCN-amplified sample CB2013. Genomic coordinates of ecDNAs in grey. Assignment of B-alleles to haplotypes by copy-number phasing in red and blue and by statistical phasing in black and green. Material from: Koche et al., Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma, Nature Genetics, published 2020, Springer Nature Limited'. Author reuse.

4.2.4 Circularised focal amplifications, but not circles in regions of balanced copy-number, show strong effect on allele-specific expression

Intrigued by the co-occurence of circles and focal amplifications we investigated the relationship between circles and transcription levels. We determined allele-specific expression at heterozygous SNPs (Section 3.1.5) to characterize expression preferences in both circularised and un-circularised regions. From 2,306,109 expressed heterozygous SNPs across all samples 7% showed allelic expression imbalance (AEI). In contrast, 95% of 495 expressed SNPs residing in focally amplified circles showed AEI, of which 99% were predominantly expressed from the circularised allele. ASE ratios for all expressed SNPs averaged to 0.6343 (95% CI 0.6342-0.6345) compared to 0.9629 (95% CI 0.9563-0.9694) for the subset of SNPs in circularised focal amplifications, indicating extreme allele-specific expression in these regions. Generally, we found RNA-seq and WGS B-allele frequencies in imbalanced genomic regions to be strongly correlated (Pearson's r=0.6154, 95% CI 0.6139-0.6168), showing a divergence from 0.5 relative to the underlying copy-number imbalance (Figure 42). To determine if circles contribute to ASE independent of the copy-number state of the underlying genomic segment we tested for preferential expression of mono-allelic circles in copy-number-balanced regions. Out of 4,193 circles with AEI in balanced CN regions we found 2,135 and 2,058 preferentially expressed from the circularised and un-circularised allele respectively (binomial test for equal probability, P = 0.24). Thus, we could not conclude that circles affect ASE independent from copy-number. These findings suggest that DNA copy-number drives allele-specific expression in imbalanced regions, including focal amplifications. Circles from focal amplifications are almost exclusively expressed from the circularised, amplified allele, likely due to strong copy-number imbalances introduced by ecDNAs. Circles outside of copy-number imbalances do not generally lead to allelic expression differences, suggesting that most circular DNA molecules are subclonal and lowly abundant or transcriptionally inactive.

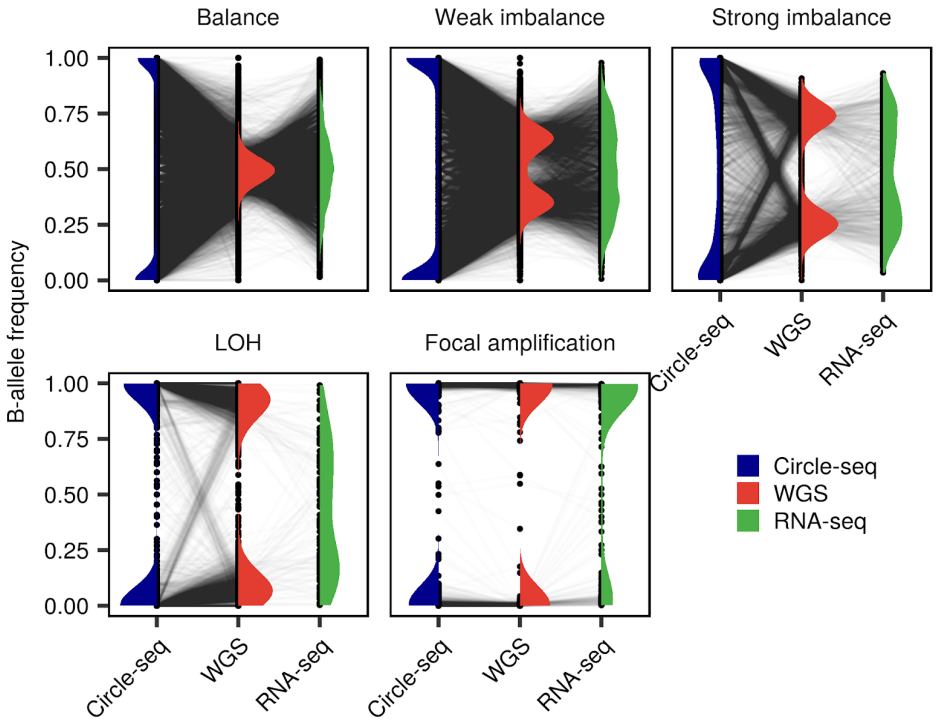


Figure 42: B-allele frequencies in Circle-seq, WGS and RNA-seq. BAF measured at the same heterozygous SNP across the three sequencing assays are connected by lines. Plots of BAF densities for each sequencing assay are shown in the respective column. Material from: Koche et al., Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma, Nature Genetics, published 2020, Springer Nature Limited'. Author reuse.

4.2.5 Multiple ecDNA-associated gene amplifications in a primary tumor

We identified multiple ecDNA calls overlapping CN amplifications in tumor CB2001 and we examined this sample in greater detail. We found 5 and 2 ec-DNA-associated amplifications on chromosome arm 2p and 1p respectively (Figure 43). In amplifications Circle-seq and RNA-seq BAFs showed strong overrepresentations of the major CN allele. Three ecDNA calls larger than 10 kb did not overlap amplifications on 1p. One ec-DNA-associated amplification was found at the MYCN locus and another overlapped genes CRIM1, FEZ2 and AC007401.2 in approximately 19 Mb distance from the MYCN locus (Figure 43a). CRIM1 is a transmembrane protein that may be subject to growth factor binding and is involved in nervous system development (Kolle et al. 2000). Recently, upregulation of CRIM1 circular RNA was found to promote nasopharyngeal carcinoma cell metastasis (Hong et al. 2020). FEZ2 is involved in axonal outgrowth (Bloom and Horvitz 1997; Fujita et al. 2004) and AC007401.2 an uncharacterized protein downstream of FEZ2.
ec-DNA-associated amplifications on 1p resided in a broader region of 11q LOH and the retained/major allele was found to be circularized (Figure 43c).

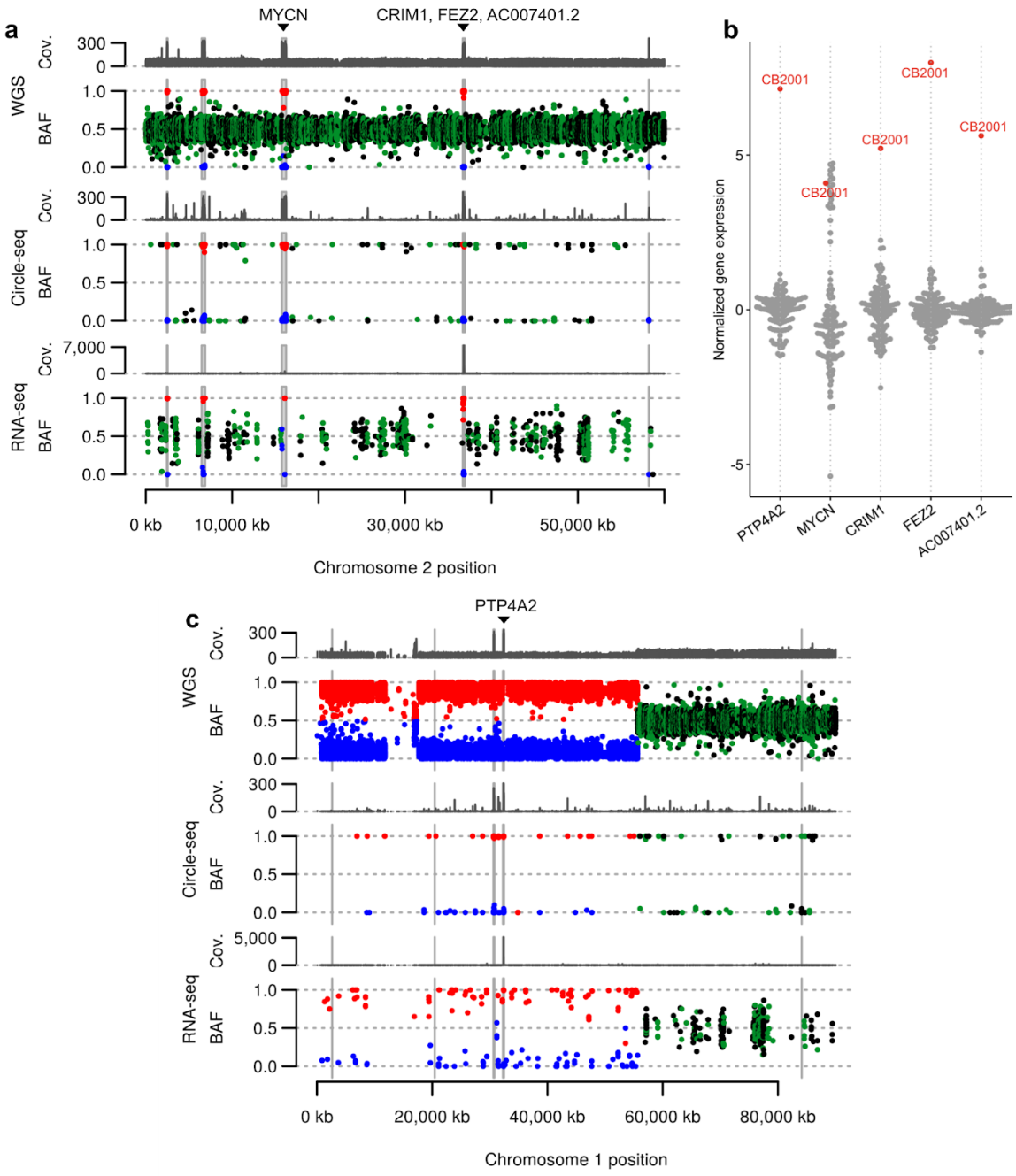


Figure 43: Extrachromosomal circular DNA-associated amplifications in tumor CB2001. Genomic coordinates of ecDNAs larger than 10 kb in grey. Assignment of B-alleles to haplotypes by copy-number phasing in red and blue and by statistical phasing in black and green. **a**, Multiple genomic regions involved in ecDNA formation including the MYCN locus on 2p. **b**, Amplification of PTP4A2 is associated with circularization of the major allele haplotype in a broader region of LOH on chromosome arm 1p. **c**, Gene expression residuals of genes overlapping ec-DNA-associated amplifications shown in (a) and (b) across 116 tumors. Sample CB2001 highlighted in red. Material (a,c) from: Koche et al., Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma, Nature Genetics, published 2020, Springer Nature Limited'. Author reuse.

One of the two amplifications on 1p overlapped the protein coding gene PTP4A2, an oncogenic phosphatase (Cates et al. 1996) that was found to be overexpressed in prostate tumor cells (Qin Wang et al. 2002). We compared RNA expression levels of ecDNA-amplified genes in CB2001 compared to other tumors and found CB2001 to be a strong expression outlier showing highest RNA levels of PTP4A1, CRIM1, FEZ2 and AC007401.2 among all tumors as well as a MYCN expression levels similar to other MNA tumors (Figure 43b).

4.3 Discussion

Previous studies on ecDNA in cancer were mainly based on cytological observations of larger double minute chromosomes and limited in their ability to detect circular DNA including smaller eccDNAs genome-wide. We here combined DNA-sequencing-based detection of circular DNA with allele-specific copy-number and ASE quantification, allowing for a broader description of circular DNAs of different sizes in terms of copy-number and gene expression of circularized alleles. Our work provides a detailed description of the interplay between circular DNA, copy-number and allelic regulation in neuroblastoma. We showed that circular DNA is mono-allelic and that somatic copy-number imbalance determines both expression imbalance and the frequency of allelic origin in smaller circular DNA outside of focal amplifications in these tumors. We do not see evidence for copy-number alterations or allelic expression differences introduced by small circular DNAs (eccDNA). In contrast, we found larger circular DNAs to be strongly associated with focal amplifications, defined by small copy-number segments of high sequencing coverage in tumor samples. Our findings confirm that large circular DNA is able to induce mono-allelic copy-number increases, that ultimately lead to an extreme expression of the circularized allele. Also, these results provide evidence for circular DNA-induced amplifications of other genes than MYCN or its neighboring genes in neuroblastoma. In fact these amplified circular DNAs may encompass both MYCN-distal regions on chromosome 2 as well as genes from different chromosomes, as we demonstrate for the protein coding genes PTP4A2, CRIM1, FEZ2, AC007401.2 which show ecDNA-associated amplification and strong overexpression in one of the tumors.

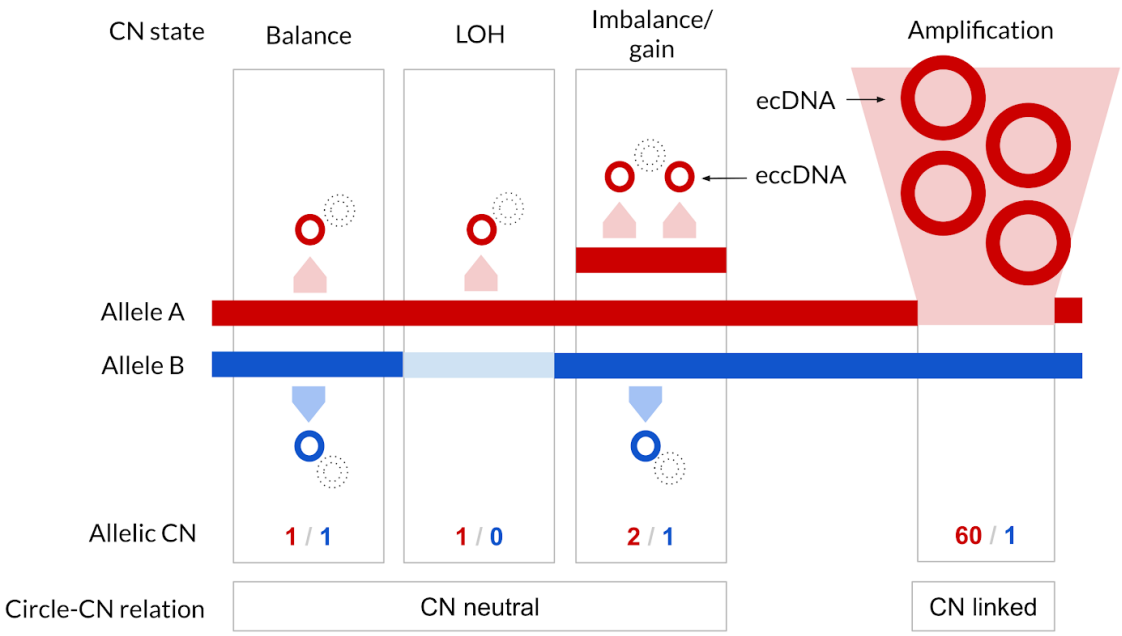


Figure 44: Circular DNA size and its relation to copy-number. Small ecDNA (eccDNA) (< 5 kb) is copy-number neutral and remains subclonal. Large ecDNA (10 kb – 2 Mb) is clonal, contains amplified DNA sequences and is therefore inherently linked to CN alterations.

Small extrachromosomal circular DNAs

We suggest that the smaller eccDNA is predominantly subclonal, because of the following observations: First, it is not associated with copy-number segments that coincide with circularized genomic regions and is thus CN neutral, meaning that there are no clonal DNA copies introduced by these small DNA molecules. And second, in small eccDNA there is generally no detectable difference in expression between the circularized and non-circularized alleles. The frequency of allelic origin of eccDNAs is driven by CN imbalance of larger CN segments. In other words, the more abundant chromosomal DNA of an allele is, the more frequently do eccDNAs from this allele occur. This finding suggests that eccDNAs are created by an almost uniform process from chromosomal DNA and the vast majority may quickly be degraded shortly after. We cannot directly distinguish tumor cell circular DNAs from those in normal cells unless they contain tumor-specific DNA lesions, which we did not examine in this work. However, it is safe to assume that a large share (if not all) of eccDNAs originate from the cancer cells fraction of the tumor sample, because eccDNA haplotype frequencies follow the patterns of somatic copy-number. Because cancer types have characteristic patterns of SCNAs, this could in part explain why chromosomal locations of eccDNA delineate groups of prostate and ovarian cancer cell lines (Dillon et al.

2015). Our findings are inline with previous reports on abundant small eccDNA in somatic tissues, which may result from deleted DNA, by-products of DNA damage (Henrik Devitt Møller et al. 2018) or replication stress (Paulsen et al. 2018). In contrast to small eccDNA, large ecDNAs are inherently linked to somatic CN, as their boundaries coincide with small CN segments. These segments are frequently amplified which supports the established view that larger ecDNAs can be clonal carriers of amplified sequences. Figure 44 shows a schematic representation of circular DNA size and its relation to copy-number based on our observations in neuroblastoma tumors.

Large extrachromosomal circular DNAs

Co-amplification of multiple distal genes on the same chromosome (Figure 43) indicates that ecDNA biogenesis of co-amplified loci could be the result of the same mutational event in the clonal history of the tumor. Infact, an analysis of structural variation in sample CB2001 revealed that the identified ecDNAs-associated amplifications on chromosome 1p and 2p are interconnected (compare Figure 43a,c with Supplementary figure 11 in Koche et al. 2020), providing evidence for the formation of chimeric circles that contain multiple co-amplified genes. Complex patterns of rearrangements complicate the mapping of exact ecDNA sequences, but together with the observations of chromosomal clusters of ecDNAs they provide growing evidence for the role of chromothripsis in the formation of larger ecDNAs (Korbel and Campbell 2013; Shoshani et al. 2020). ecDNAs are not necessarily stable during tumor cell evolution and already early studies reported re-integration of these molecules into homogeneously staining regions on different chromosomes (Kohl et al. 1983; Carroll et al. 1988), implying that ecDNA is a source of genomic instability. In terms of SCNAs, we found MNA tumors to be genomically more stable than other neuroblastomas tumors (Figure 20). However, as MYCN amplifications frequently reside on ecDNAs their reintegration into distal genomic loci could lead to genetic heterogeneity. Indeed, neuroblastoma tumors with circle-associated rearrangements have a worse prognosis, which holds true even within the group of MNA tumors (Koche et al. 2020, Figure 4d-e) but not for rearrangements in general (Koche et al. 2020, Supplementary figure 12a-c). This suggests that ecDNA-induced genomic instability could be an important source of clinical heterogeneity between high-risk tumors of the MNA subgroup (Koche et al. 2020).

Copy-number-independent gene regulation by circular DNAs

Apart from the regulatory potential of ecDNAs by DNA copy-number dosage that we have examined here, ecDNA can also result in regulatory changes for genes at chromosomal

integration sites and for genes on the circularized DNA itself. Our allele-specific ecDNA analysis (Section 4.1.3) provided evidence for the re-integration of a MYCN amplicon-derived ecDNA sequence into chromosome 13 in tumor sample CB2013, as together with Koche et al. we could show that the genotype of heterozygous SNP rs13028343 was identical between the amplified and integrated sequence; and this ecDNA integration disrupted the coding sequence of tumor suppressor DCLK1 (Koche et al. 2020, figure 3d and 4b). Similarly, integration of ecDNAs in non-coding regions can facilitate the hijacking of regulatory elements. Evidence for this is provided by ecDNA-associated rearrangements of a MYCN-amplicon sequence upstream of TERT in CB2027 with high expression of TERT and its neighboring gene SLC6A18 (Koche et al. 2020, figure 4c). ecDNA of several hundert kb to a few megabases in size are expected to contain a substantial proportion of non-coding sequences from their chromosomal site of origin. These intergenic sequences contain CREs that are known to control the transcriptional program on the “linear” chromosome and are associated with accessible chromatin (Section 2.2.1). However, compared to chromosomal DNA, ecDNAs were found to contain even higher degrees of accessible chromatin and could facilitate long-range chromatin interactions within circularized structures (S. Wu et al. 2019). Interestingly, on ecDNAs regulatory elements were found to be preferentially co-amplified with oncogenes like EGFR in glioblastoma and MYCN in neuroblastoma, indicating that some enhancer elements on ecDNAs remain functional, were sometimes hijacked from distal chromosomal regions and contributed to the formation of de-novo topological interactions that do not form in the context of chromosomal DNA (A. R. Morton et al. 2019; Helmsauer et al. 2020).

Summary

In conclusion our results show that large but not small circular DNA is associated with genomic copy-number amplification, leading to strong expression exclusively from the circularized allele. Co-amplification of distal genomic regions indicates that mosaics of genomic sequences are co-amplified on large ecDNA, which strengthens the hypothesis of chromothripsis being the mechanism of large ecDNA formation. Smaller eccDNA, that is also present in healthy somatic tissue, is a class of highly abundant circular DNAs that is distinct from its larger amplification-associated counterpart. Small eccDNAs do generally not introduce clonal copy-number alterations. Their frequency of allelic origin is determined by larger underlying copy-number segments, which prioritizes transcriptional or DNA replication stress as causes of eccDNA formation. If small eccDNAs can give rise to larger circular DNAs that may manifest themselves clonally is still an open question. Given the smaller size

of eccDNA it is less likely that they contain complex CRE arrangements that are topological associating as described for ecDNAs. However, we here only characterize dominant characteristics of these two classes of circular DNAs and individual small eccDNAs may still be transcribed and manifest themselves as clonal copy-number alterations.

5 Germline cis-regulatory variation

The previous chapters focused on somatic variation and its impact on gene expression. In the analysis in chapter 3 I showed that after somatic CN, germline variation is overall the second strongest local genetic contributor to ASE and total expression variance of all factors considered (Section 3.2.3). In this chapter, I will focus on the regulatory role of germline variants. To identify SNPs and genes involved in cis-regulation in neuroblastoma, genotypes will be associated with gene expression and ASE by cis-eQTL and cis-aseQTL mapping respectively. These analyses identify genes of heritable gene expression as well as SNPs linked to the two quantitative traits. A comparison of the mapping strategies reveals biases in cis-aseQTL mapping that underly discrepancies in the results obtained. I will here describe and discuss these biases. To prioritize functional variation, epigenetic readouts in neuroblastoma cell line SH-SY5Y will be integrated with eQTL results from the 116 primary tumors. I will also characterize potential cis-regulatory effects of SNPs at loci that were found to be associated with neuroblastoma susceptibility by a previously reported GWAS.

Contributions to this chapter

Alignments of normal WGS, tumor WGS, tumor RNA-seq were created by the Core Unit Bioinformatics (CUBI) of the Berlin Institute of Health (Berlin, Germany) under supervision of Dr. Dieter Beule. Remo Monti (AG Ohler, MDC, Berlin, Germany) integrated the fastlmm and PEER methods into the data processing pipeline that were used to map cis-eQTL and cis-aseQTLs. Dr. Dubravka Vucicevic (AG Ohler, MDC, Berlin, Germany) prepared ATAC-seq libraries and Dr. Scott Lacadie (AG Ohler, MDC, Berlin, Germany) generated peak calls and signal tracks from ATAC-seq.

5.1 Methods

5.1.1 cis-QTL association testing

For eQTL analysis the SNP genotypes called in 116 WGS samples of normal tissue were pooled and filtered. Only SNPs with a minor allele frequency of 5% and at least 10% genotyped samples in the cohort were retained. Htseq count (Anders, Pyl, and Huber 2015) was used to count reads from RNA-seq data of tumor samples in the union of all exons per gene based on the Ensembl 75 gene annotation (Section 3.1.3). Raw RNA gene counts

were normalized by library depth per sample and transformed to variance-stabilized counts by DESeq2 (Love, Huber, and Anders 2014). Only protein-coding genes on chromosomes 1-22 with at least 10 counts in 90% of the samples were considered. In total 13,903 genes were included in the analysis. Variance-stabilized counts were centered and strong outlier samples, defined as normalized count values exceeding 3 times the standard deviation of all normalized counts per gene, were removed. To estimate the expression variability between samples we applied probabilistic estimation of expression residuals (PEER) (Stegle et al. 2012) to derive 10 factors from the normalized counts. We took these factors as representatives for global expression differences, that are likely not associated with cis-regulatory effects and incorporated them as covariates in the association test described below. Genotypes of SNPs in a cis-window of 500kb upstream and downstream of annotated gene coordinates were associated with the gene's quantiative trait (Ffigure 45). SNPs were associated with quantitative traits by FastLMM (Christoph Lippert et al. 2011), version 0.2.23 in single SNP mode. FastLMM uses a linear mixed model in a regression of the number of alternative alleles on quantitative phenotypes controlling for given covariates. We combined gene- and sample-specific covariates individually in each test. The somatic gene copy-number was calculated as the average total copy-number in gene intervals and used as the only gene-specific covariate. Sex, cohort membership, tumor purity, tumor ploidy and the 10 PEER factors were incorporated as sample-specific covariates. Each association test was controlled by the matching set of sample and sample-gene-specific covariates for a given set of gene associations.

Similarly, we conducted an aseQTL analysis to associate SNP genotypes in the same cis-window as described above to the ASE quantitative trait (Section 3.1.5). Cis-effect SNPs of homozygous genotype could both result in high and low total expression but are not expected to induce ASE, because here both haplotypes share the same regulatory effect (Figure 9). To map genotypes to ASE we collapsed homozygoutes of both reference and alternative allele to a common genotype state (homozygous) and contrasted it with the heterozygous state. aseQTL mapping was then carried out using FastLMM (same version as above) using sex, cohort membership, tumor purity, tumor ploidy as global covariates and the copy-number ratio as local gene-level covariates. Because the ASE phenotype should reflect cis-effects specifically and is well controlled for trans effects and global RNA count biases between samples we did not control for PEER covariates in aseQTL association test.

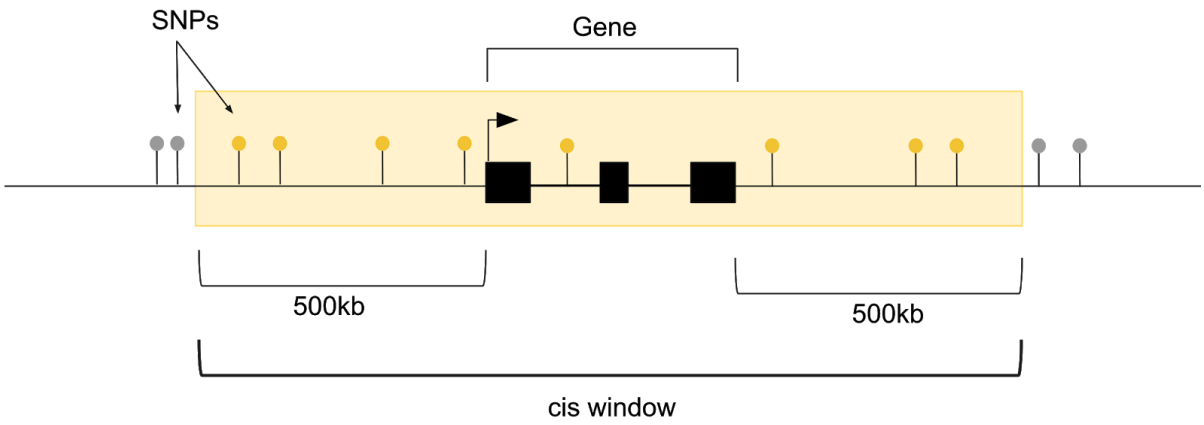


Figure 45: SNPs in cis-window relative to gene coordinates are associated with quantitative trait. A cis-window spanning 500kb upstream and downstream from the annotated coordinates was defined for each gene tested. SNPs inside this window (yellow circles) were included in the association test between the individual SNP genotypes and the quantitative trait of the gene.

We defined eQTL- and aseQTL genes as those genes in which we detect genome-wide significant associations between cis-SNPs and the respetive quantitative trait. To determine these genes we applied a hierarchical p-value correction procedure after removing SNPs with highly correlated genotypes resulting in one p-value per gene (Ffigure 46). SNPs with highly correlated genotypes were removed by a linkage disequilibrium (LD) filter as implemented in PLINK (Purcell et al. 2007) (version 1.9) in indep-pairwise mode with window size 50, step size 5 and r-squared threshold 0.9. In the first correction step, the remaining SNP-gene-association p-values were adjusted by the Bonferroni method for all SNPs in the gene's cis-window. The minimum Bonferroni adjusted p-value per gene was then considered as a representative nominal gene-level p-value. In a second step, these p-values were adjusted for multiple-testing on genome-level by the Benjamini-Hochberg method. QTL genes were defined as those genes with a SNP association significant at FDR 0.05 among genome-level p-values.

Lead eQTL and aseQTL SNPs were identified for each eQTL and aseQTL gene by considering all association tests in the respective cis-windows. SNPs of association tests with the minimum nominal p-value as determined by FastLMM were marked as lead eQTLs or lead aseQTLs depending on the underlying test strategy. We identified SNPs in strong LD ($r^2 > 0.9$) with lead eQTLs (here referred to as "LD SNPs") using PLINK (version 1.9) and parameters --show-tags and --tag-r2 0.9. By default this command scans for SNPs exceeding the defined correlation threshold in a 250 kb window around given target SNPs. We annotated promoter-proximal SNP tests by their distance to the corresponding gene's

TSS. We first defined the TSS as the gene's start and end position (based on gene annotations in Ensembl version 75) on the positive and negative strand respectively. Then, the TSS distance of a SNP and its corresponding gene was determined as $d_{TSS} = s_{gene}(p_{SNP} - p_{TSS})$, where p_{SNP} is the position of the SNP, p_{TSS} the position of the TSS and s_{gene} is defined as +1 and -1 for genes on the positive and negative strand respectively. SNP tests, which satisfied $-2000 < d_{TSS} < 500$, were marked as promoter-proximal in relation to the tested gene.

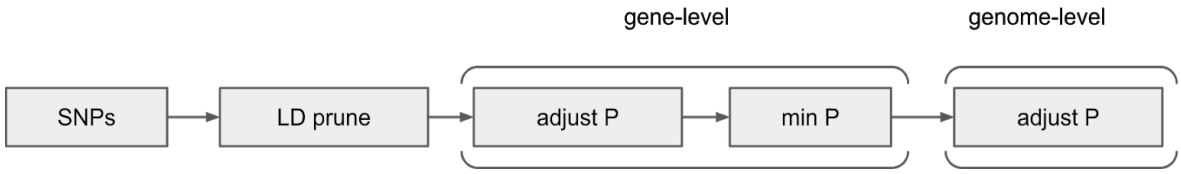


Figure 46: p-value adjustment procedure to determine QTL genes. Significant QTL genes were identified by a two-step p-value correction procedure after pruning SNPs in close linkage disequilibrium (LD).

5.1.2 ATAC-seq and H3K27ac-ChIP analysis

To prioritize cis-regulatory SNPs we integrated QTL mapping results with epigenetic readouts in neuroblastoma cell line SH-SY5Y. We derived peaks and read coverage signals from ATAC-seq and H3K27ac ChIP-seq to identify genomic regions that are accessible and therefore potentially harbor CREs. Reads of H3K27ac ChIP-seq in cell line SH-SY5Y were obtained from the NCBI sequence read archive[16] accession SRR3363257 (Henrich et al. 2016). Raw reads were mapped to the human reference genome GRCh38/hg38 by bowtie2 (Langmead et al. 2009) version 2.3.4.3 in single-end mode with default parameters. Fragment lengths were estimated by MACS2 (Feng, Liu, and Zhang 2011) version 2.1.1.20160309. Peaks were called using JAMM (Ibrahim, Lacadie, and Ohler 2015) version 1.0.8 with parameters -e auto -b 100 -t single and -f set to the estimated fragment length. BED files were generated from coordinates of ChIP-seq read alignments and extended to the estimated fragment length. BED files with extended read mapping coordinates of H3K27ac ChIP-seq and BED files of corresponding peak coordinates were translated to GRCh37/hg19 coordinates using R/Bioconductor package rtracklayer (Lawrence, Gentleman, and Carey 2009) version 1.4.6. H3K27ac ChIP-seq signals were determined by coverage based on the generated BED files.

[16] https://www.ncbi.nlm.nih.gov/sra, accessed 18 Mar 2021

Library preparation for ATAC-seq in neuroblastoma cell line SH-SY5Y was performed on 100,000 cells according to established protocols (Buenrostro et al. 2015) with the following modifications: transposition time was increased from 30 min to 1 h and the cell pellets were taken directly to the transposition reaction omitting the lysis step as previously described (Karabacak Calviello et al. 2019). For all samples 12 PCR cycles were performed and the libraries were sequenced (2x75nt) on a NextSeq 500/550 using a HighOutput v2 Kit for 150 cycles (Illumina #FC-404-2002, discontinued). Sequencing adapters were trimmed from raw reads using FLEXBAR (Dodt et al. 2012). Processed reads were then aligned to the GRCh37/hg19 reference by bowtie2 with parameters -p 4 -X 1500 --no-discordant. Alignments were filtered by samtools[17]to keep uniquely mapping reads only and duplicates were removed by picard-tools[18] 1.90. Remaining mapping coordinates were reduced to the reads' 5' end and extended to a fixed length of 38 bp using bedtools (Quinlan and Hall 2010). Peaks in the modified alignments were called by JAMM 1.0.7.5 with parameters -f 38 -b 100 -e auto -p 4.

We quantified H3K27ac and ATAC signals at SNP positions by counting reads in broader regions surrounding the SNPs. To that end we extended single base pair positions of QTL SNPs by ±1000 bp and ±200 bp and counted overlapping reads from H3K27ac ChIP-seq for ATAC-seq respectively. Overlaps between extended SNP intervals and coordinates of aligned fragments (modified as described above) were counted by the countOverlaps method of R/Bioconductor package GenomicRanges version 1.38 with parameters type="any" and ignore.strand=TRUE. SNPs in ATAC-seq and H3K27ac ChIP-seq peaks were determined by overlap of genomic coordinates using the method overlapsAny from package GenomicRanges version 1.38.

5.1.3 Enrichment test of ATAC-seq and H3K27ac ChiP-seq signal

To test for enrichment of lead eQTLs and LD SNPs in ATAC-seq and H3K27ac ChIP-seq signals a permutation-based enrichment test was conducted. SNPs in cis-windows of eQTL genes were ordered by decreasing signal strength determined at windows around SNP positions (Section 5.1.2). The score $s_i = 1/k$ was assigned to a SNP i if it was a member the group tested (e.g. lead eQTL or any of lead eQTL and LD SNP) and $s_i = -1/(n-k)$ to SNPs outside of the group, where k is the number of SNPs within the group and n the total number of SNPs. The cumulative sum m_i along the list of signal-ordered SNPs $t^* = abs(max(m_i))$ was

[17] http://www.htslib.org/, accessed 18 Mar 2021
[18] http://broadinstitute.github.io/picard/, accessed 18 Mar 2021

calculated. Then the procedure was repeated 10,000 times by permuting the group assignments of the SNPs and for each permutation *j* the absolute cumulative sum of scores t_j was determined as described above. The empirical p-value $p = n_g / 10{,}000$ was calculated, where n_g is the number of permutations for which $t_j \geq t^*$. If n_g was found to be equal 0, the p-value was reported as the upper bound $p < 1 / n_p$, where n_p is the number of permutations. We used 10,000 permutations and the upper bound was $p < 1.0 \times 10^{-4}$ accordingly.

5.1.4 Test for deviation from Hardy-Weinberg principle

In contrast to eQTL mapping, where generally all samples are available for association tests, aseQTL is often limited to a subset of samples for a given gene. This is because not all samples may be informative for ASE in that gene. We examined genotype biases in aseQTL mapping that may occur by considering non-random subsets of samples. To investigate differences between eQTL and aseQTL mapping results we determined biases in lead eQTL genotypes specifically. The frequency of expected genotype distributions in random-mating populations in the absence of selection (or mutation) can be inferred from allele frequencies by the Hardy-Weinberg principle (HWP). The HWP states that in the case of two alleles A and B and corresponding population allele frequencies *p* and $q = 1-p$ the genotypes AA, AB and BB are expected to be observed with frequencies p^2, $2pq$ and q^2 respectively (Stern 1943). To identify biases in genotypes we first determined allele frequencies *p* and *q* at lead eQTLs and the expected allele counts of homozygous A (E_{AA}), heterozygous (E_{AB}) and homozygous B (E_{BB}) genotypes. We then compared expected to observed genotype counts as available to eQTL- and aseQTL mapping using a Chi-squared test (Emigh 1980). The corresponding test statistic is given by

$$\chi^2 = \sum_{g \in G} \frac{(n_g - E_g)^2}{E_g} \quad , \tag{5}$$

where *G* is the set of possible genotypes {AA, AB, BB}, n_g the observed number of samples for genotype *g* and E_g the expected number of samples for that genotype. According to the HWP the expected number of samples for the three possible genotypes of a bi-allelic trait is given by $E_{AA} = np^2$, $E_{AB} = n2pq$ and $E_{BB} = nq^2$. We identified deviation from HWP in ASE-informative subsets of lead eQTL genotypes by rejection of the null hypothesis of independence determined by the χ^2 statistic controlling for multiple testing burden at FDR 0.05 (Benjamini-Hochberg).

5.2 Results

5.2.1 Expression and allele-specific expression quantitative trait loci

We tested associations between SNP genotypes and quantitative gene expression traits to identify genes with genetic variability in cis-regulation in neuroblastoma primary tumors and to identify candidate SNPs involved in this form of regulation. To that end genotypes of SNPs, as obtained by our analysis described in section 3.1.4 in a local genetic environment defined by a cis-window from -500 kb upstream from the gene start to +500 kb downstream from the gene end, were associated with two different gene expression traits: Total gene expression (Section 3.1.3) was used in eQTL- and ASE (Section 3.1.5) in aseQTL mapping. Quantitative traits were controlled for covariates obtained from the clinical annotation, as well as cohort membership, global gene expression covariates and local copy-number effects. To determine genes with genome-wide significant genetic variability in cis-regulation we considered a subset of QTL associations of SNPs in weaker LD and applied a hierarchical p-value correction strategy, that controlled for gene-level discoveries at FDR 0.05. Section 5.1.1 gives a detailed description of the applied QTL mapping and the statistical approach.

In total 24,527,007 and 23,026,537 gene-SNP combinations were tested for association in eQTL and aseQTL analysis respectively. On average 1775 (SD ± 815) SNPs were considered in cis-windows of eQTL mapping. In aseQTL analysis an average of 1820 (SD ± 839) SNPs were considered. To identify genes with significant expression heritability by cis-regulation highly correlated SNP genotypes ($r^2 > 0.9$) were removed before gene-level statistics were calculated. In this gene-level analysis, an average of 540 (SD ± 290) and 554 (SD ± 295) cis SNPs per gene were considered in eQTL and aseQTL analysis respectively. Our approach identified 163 genes with total gene expression to be significantly associated with SNP genotypes in cis (eQTL genes). The gene-level aseQTL approach revealed 24 genes with significant associations between ASE and heterozyousity of SNPs in cis (aseQTL genes). For each eQTL and aseQTL gene identified we determined the SNP with strongest association to the respective trait in the complete set of cis-window SNPs. These SNPs were defined as lead eQTLs and lead aseQTLs respectively. A median of 1 (1–37) lead eQTLs and 1 (1–28) lead aseQTLs were determined. To examine how QTL associations were distributed relative to gene coordinates we calculated the distance of each associated SNP to the gene's TSS. Here, the TSS position was defined as the 5' end of the gene according

to the Ensembl annotation[19]. Here, negative and positive TSS distances indicate upstream and downstream locations relative to the TSS respectively. We found the highest density of strong SNP associations was located close to the gene's TSS and lead eQTLs density peaked around the TSS with the majority being located downstream (Figure 47a). Multiple lead eQTLs of identical association p-values spanned distances of up to 112 kb, suggesting long range LD between cis-regulatory variants and individual SNPs. Alternative allele counts at lead eQTL were associated with increased or attenuated expression of eQTL genes across the cohort (Figure 47b). Figure 47 shows eQTL association p-values relative to the gene's TSS and the lead eQTL genotype effect on expression in the top 30 eQTL genes identified. The aseQTL mapping showed similar results and the majority of identified lead aseQTLs was located downstream of the TSS. However, lead eQTL density at the TSS was less pronounced. The highest density peak was dominated by a larger cluster of 28 lead aseQTLs of identical association p-values 166–214 kb downstream of the TSS of gene HNRNPH3 (Supplementary figure 9a). Among the 24 aseQTL genes the strongest median ASE was consistently observed for heterozygous genotypes of lead aseQTLs (Supplementary figure 9b). Supplementary table 17 lists results of eQTL and aseQTL association tests.

[19] Ensembl release 75, http://www.ensembl.org/, accessed 18 Mar 2021

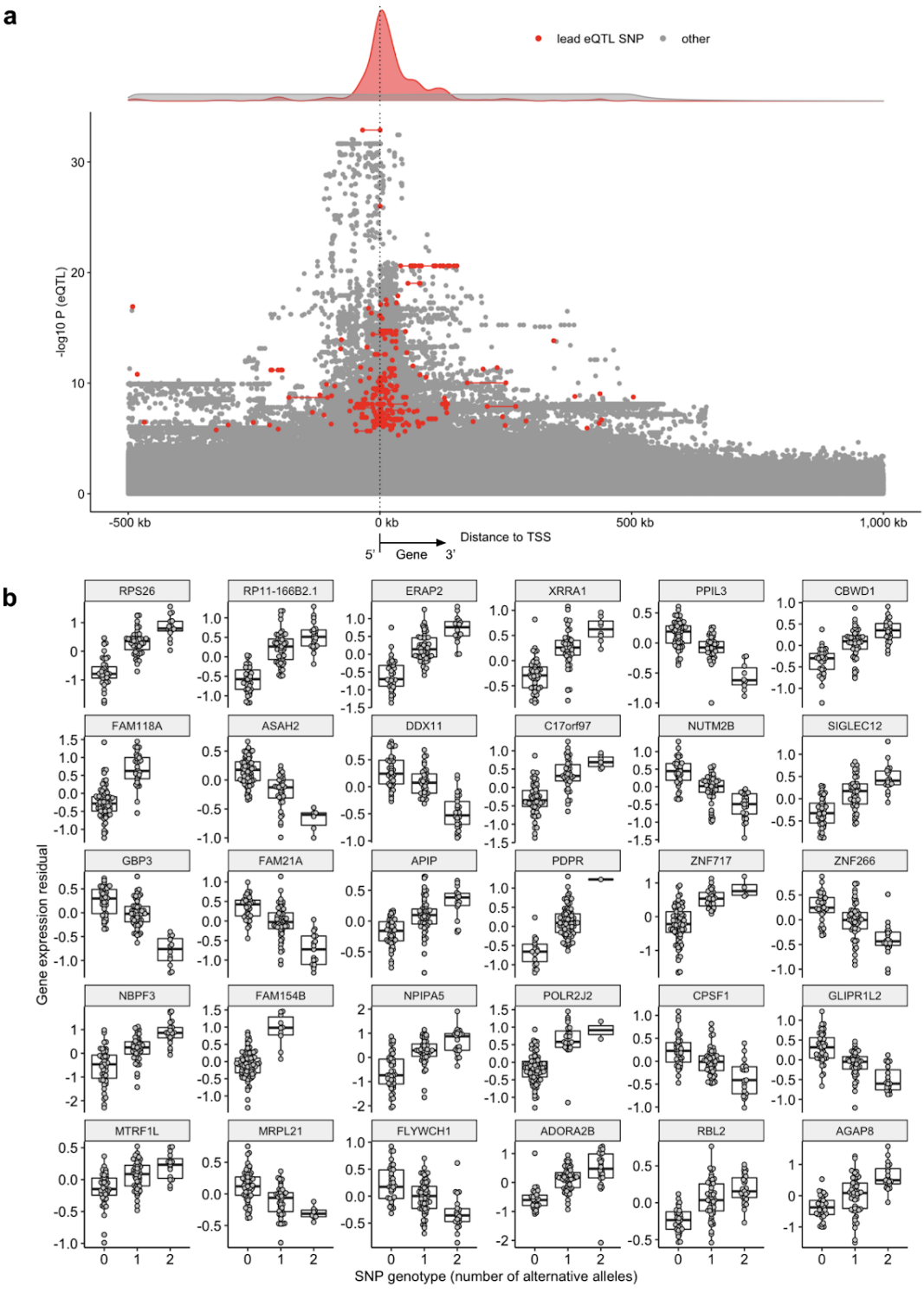


Figure 47: Expression quantitative trait loci associations. **a**, eQTL associations by distance from gene transcription start site (TSS). Top: Density of TSS distance. Bottom: eQTL association p-value by TSS distance. Association tests of SNPs > 1,000 kb TSS distance are not shown. If there are multiple lead eQTL SNPs then they are connected by a red line. Gene's 5' to 3' direction indicated below the plot. **b**, Expression by genotype of lead eQTL SNP for 30 strongest associated eQTL genes. For genes with multiple lead eQTL SNPs and arbitrary one if selected.

We noted the remarked difference in the number of eQTL and aseQTL genes identified, with far fewer significant aseQTL than eQTL genes (24 to 163). Generally, we expect that aseQTL analysis is more sensitive to detect cis-regulation of genes that are strongly regulated by trans effect. Strong trans effects could obscure associations between expression and SNP genotypes in eQTL analysis. However, we would then still expect that identified eQTL genes harbor substantial effects in aseQTL mapping, because the eQTL regression is based on phenotypic differences explained by allele counts (compare Figure 9a and b). Nevertheless, after adjusting for multiple testing correction ACCS is the only gene we found to be significant in both QTL mapping strategies. Comparison of p-values between lead eQTL and lead aseQTL associations revealed that many observed effects were exclusive to total expression and ASE mapping respectively (Supplementary figure 10). We speculated that a low number of ASE informative samples in an eQTL gene would make associations with the ASE phenotype more difficult, which is one possible explanation for high aseQTL p-values in significant eQTL associations. In addition, we observed marked differences in the distribution of genotypes of lead SNPs in eQTL and aseQTL genes (compare Figure 47b and Supplementary figure 9b) and hypothesized that these differences could be a cause for missing aseQTL associations at eQTL loci. To understand if the subset of ASE informative samples skews the distribution of available genotypes we tested for deviation from the Hardy-Weinberg principle (HWP). To this end we selected all eQTL lead SNPs and tested (1) HWP deviation across all genotypes (as visible to eQTL mapping) and (2) HWP deviation across genotypes in ASE informative samples (as visible to aseQTL mapping). HWP deviation was determined by Chi-squared test of observed and expected genotype frequencies and FDR 0.05 (Benjamini-Hochberg) (Section 5.1.4). We found that genotypes in ASE informative samples deviated significantly from HWT in 40 out of 160 lead eQTL SNPs considered and that heterozygous genotypes were frequently overrepresented, while homozygous genotypes tended to be underrepresented (Supplementary figure 11b). To investigate which factors influence diverging eQTL and aseQTL results we then modeled the differences in -log10 transformed p-values by coverage at hetSNPs (log), the number of ASE informative samples and the imbalance between heterozygous and homozygous genotypes in ASE informative samples for each lead eQTL SNP. The genotype imbalance was defined as *het* / (*het* + *hom*), where *het* is the number of heterozygotes, and *hom* the number of homozygotes of ASE informative samples. We find all factors to be highly significant (both hetSNPs coverage and genotype imbalance $P < 2 \times 10^{-16}$ [F-test] and number of ASE informative samples $P = 1.82 \times 10^{-9}$ [F-test]). Estimated coefficients showed a positive effect of genotype imbalance and a negative effect of ASE

SNP coverage, indicating that higher ASE SNP coverage and fewer heterozygotes increase agreement between eQTL and aseQTL associations. Genotype imbalance explained 40.9%, ASE SNP coverage 8.2% and the number of ASE informative samples 2.8% of variance in p-value differences. 48% of variance in p-value differences could not be explained by our model.

These results suggest that differences between eQTL and aseQTL mapping are strongly influenced by genotype skews as a result of subsampling to ASE informative samples and moderately influenced by ASE SNP coverage and the number of ASE informative samples. Based on our findings we can assume that these biases prevent aseQTL mapping to detect cis-regulatory SNPs at specific loci, at which these effects might still be detected by eQTL analysis. The reason for these skews remains to be investigated and therefore it is hard to directly compare results of the two mapping strategies at this point. eQTL mapping is generally not affected by subsampling skews and we therefore decided to continue the systematic analysis of cis-regulation based on eQTL mapping results. However, note that the considerations above do not suggest that significant aseQTL associations are the result of biases. We can assume that aseQTL genes are subject to cis-regulation and we will still present and discuss selected results from this analysis in the course of this chapter.

5.2.2 Prioritizing cis-regulatory SNPs

eQTL associations may reflect both functional effects of cis-regulatory SNPs as well as linkage desequilibrium between non-functional SNPs and those involved in cis-regulation. Because SNPs in promoter and enhancer regions are potential genetic regulators of transcription factor binding, they are good candidates for genetic cis-regulators. In contrast, those SNPs not found to be associated with the quantitative trait are less likely to be involved in cis-regulation in the observed tissue or disease context. We aimed to prioritize functional eQTL SNPs by the result of our association test, LD structure and SNP overlap with epigenetic marks indicative of regulatory elements. To this end we identified SNPs in strong LD with our lead eQTL SNPs and investigated chromatin accessibility and histone 3 lysine 27 acetylation (H3K27ac) at their genomic locations. Because histone modification and chromatin accessibility assays are not available for the primary tumor tissues which underlies our eQTL mapping results, we made use of epigenetic data collected in the neuroblastoma cell line SH-SY5Y instead. We assumed that functional SNPs are not limited to the strongest associated SNPs for the follwing reasons: First, multiple SNPs could contribute to cis-regulation with different effect sizes. And second, the combined effects of

multiple functional SNPs in weaker LD could increase the association signal of a non-functional SNPs in stronger LD with both functional SNPs. To address this, we included SNPs in strong LD ($r^2 > 0.9$) with lead eQTLs in our search space for putative functional variation. We overlapped ATAC-seq and H3K27ac ChiP-seq peaks with all SNPs tested for eQTL associations and additionally determined ATAC-seq and H3K27ac ChiP-seq signals at their location by read counts in the respective alignment around SNP coordinates (Section 5.1.2). To estimate if eQTL associations are enriched in cis-regulatory elements at eQTL gene loci we examined if our selected SNPs (lead eQTLs and their LD SNPs) are overrepresented in ATAC-seq and H3K27ac ChiP-seq peaks respectively. Status “eQTL (+)” was assigned to all lead eQTLs and LD SNPs. Status “ATAC (+)” to all SNPs overlapping ATAC-seq peaks. Then association of these states was determined by a Chi-squared test.

We found strong enrichment of ATAC peaks in lead eQTLs ($P = 7.41 \times 10^{-14}$, Pearson's Chi-squared test) and also substantial enrichment in the extended eQTL set consisting of both lead eQTLs and LD SNPs ($P = 7.42 \times 10^{-7}$, Pearson's Chi-squared test) (Figure 48). We have confirmed that strong eQTL associations are prevalent close to the associated gene’s TSS (Figure 47a). To examine if also distal lead eQTLs and LD SNPs are enriched in epigenetic marks of regulatory elements we repeated the enrichment test, but this time excluded promoter-proximal SNPs, defined as those SNPs within -2000 to +500 bp from the annotated gene start. We find distal lead eQTLs ($P = 2.64 \times 10^{-3}$, Pearson's Chi-squared test) to be significantly enriched in ATAC-seq peaks, but we could not detect an enrichment considering both distal lead eQTLS and LD SNPs ($P = 0.27$, Pearson's Chi-squared test) (Figure 48b). We also determined if lead eQTLs and LD SNPs are enriched in ATAC-seq signal around SNP coordinates. To this end we ranked all SNPs tested for eQTL association in cis windows of eQTL genes by their normalized ATAC-seq signal and determined an enrichment p-value by permutation testing (Section 5.1.3). We find ATAC-seq signal enriched in lead eQTLs ($P = 0.0051$, permutation test) as well as LD SNPs ($P < 1.00 \times 10^{-4}$) (Figure 48c,e). The test was repeated for the subset of distal SNPs as defined above. Here, lead QTLs and LD SNPs were enriched in ATAC-seq signal ($P < 1.00 \times 10^{-4}$, permutation test). However, we were not able to detect significant enrichment of ATAC-seq signal in distal lead eQTLs ($P = 0.174$, permutation test) (Figure 48d,f).

We also conducted enrichment tests for H3K27ac ChIP-seq peaks and signals. H3K27ac ChIP-seq peaks were strongly enriched in lead eQTLs ($P = 1.24 \times 10^{-10}$, Pearson's Chi-squared test) as well as distal lead eQTLs ($P = 6.61 \times 10^{-4}$, Pearson's Chi-squared test),

but we did not detect significant enrichment in lead eQTLs and LD SNPs (P = 0.51, Pearson's Chi-squared test) or distal lead eQTLs and LD SNPs (P=0.07, Pearson's Chi-squared test). We found H3K27ac signal to be significantly associated with lead eQTLs, LD SNPs and distal LD SNPs (all $P < 1.00 \times 10^{-4}$, permutation test) and distal lead eQTLs ($P = 6.0 \times 10^{-4}$, permutation test). Supplementary figure 13 shows results of eQTL enrichment tests of H3K27ac ChIP-seq peaks and signals.

Our results show that both ATAC and H3K27ac peaks are overrepresented in promoter-proximal lead eQTLs. While lead eQTLs together with their LD SNPs were enriched in ATAC peaks, we did not observe a similar enrichment in H3K27ac peaks, indicating that SNPs linked to lead eQTLs are often found outside this chromatin mark. We do find an enrichment of H3K27ac signals for this extended set of SNPs, but the association was mainly driven by LD SNPs harboring medium-ranked H3K27ac signals (Supplementary figure 13e,f), indicating that either functional SNPs do not lie in regions of strongest H3K27ac signal or inclusion of LD SNPs enriches for non-functional SNPs with medium-ranked H3K27ac signal. An effect we do not observe for enrichment of ATAC signal, where the enrichment score consistently increases in approximately the first quarter of SNPs with highest ATAC rank (Figure 48e,f). Binding of transcription factors (except for pioneering factors) is expected to occur predominantly in open chromatin regions (Section 2.2.1) so ATAC-seq peaks are expected to demarcate regions of possible variation affecting TF binding. Based on these consideration and the discrepancies between ATAC-seq and H3K27ac ChIP-seq signals above we decided to prioritize candidate cis-regulatory SNPs by overlap with ATAC peaks and merely report overlap with H3K27ac peaks as additional evidence. All lead eQTLs and their LD SNPs were included as potential functional SNPs and only those overlapping ATAC peaks were considered.

We find 28 of 388 lead eQTLs (7.22%) and 108 of 4,316 LD SNPs (2.5%) to overlap ATAC peaks. These 136 SNPs were considered cis-regulatory candidate SNPs. 41 of the 136 selected candidate SNPs (30.15%) were located proximal to the promoter of their respective gene, while 95 (69.85%) were distal. 41 (30.15%) of the candidate cis-regulatory SNPs were promoter proximal (within -2000 bp to +500 bp of the corresponding gene’s TSS) and 95 (69.85%) were distal, meaning they did not fall within this region.

We find that 54 (39.71%) of prioritized candidate cis-regulatory SNPs also overlapped H3K27ac peaks. From in total 163 eQTL genes identified 59 (36.2%) harbored at least one

candidate cis-regulatory SNP. Of these genes 29 (21.32%) harbored a candidate SNP at the promoter, while the remaining 39 (28.68%) harbored distal candidate SNPs exclusively. Figure 49 shows TSS distance and eQTL association p-value of prioritized candidate cis-regulatory SNPs and indicates which SNPs overlap H3K27ac peaks. Density of H3K27ac overlapping candidate SNPs showed remarked concentration at the TSS, indicating that lead eQTLs and LD SNPs in ATAC peaks have strongest overlap with H3K27ac peaks around the TSS (Figure 49a). However, those candidate regulatory SNPs that additionally overlapped H3K27ac peaks showed higher density further downstream of the TSS compared to SNPs in ATAC peaks only (Figure 49b), indicating that a prioritization strategy that would require the presence of both chromatin features would likely miss regulatory SNPs upstream of the TSS. Supplementary table 18 lists all SNP identifiers of prioritized candidate cis-regulatory SNPs for identified eQTL genes and their respective eQTL association p-value, TSS distance and H3K27ac ChIP-seq peak overlap.

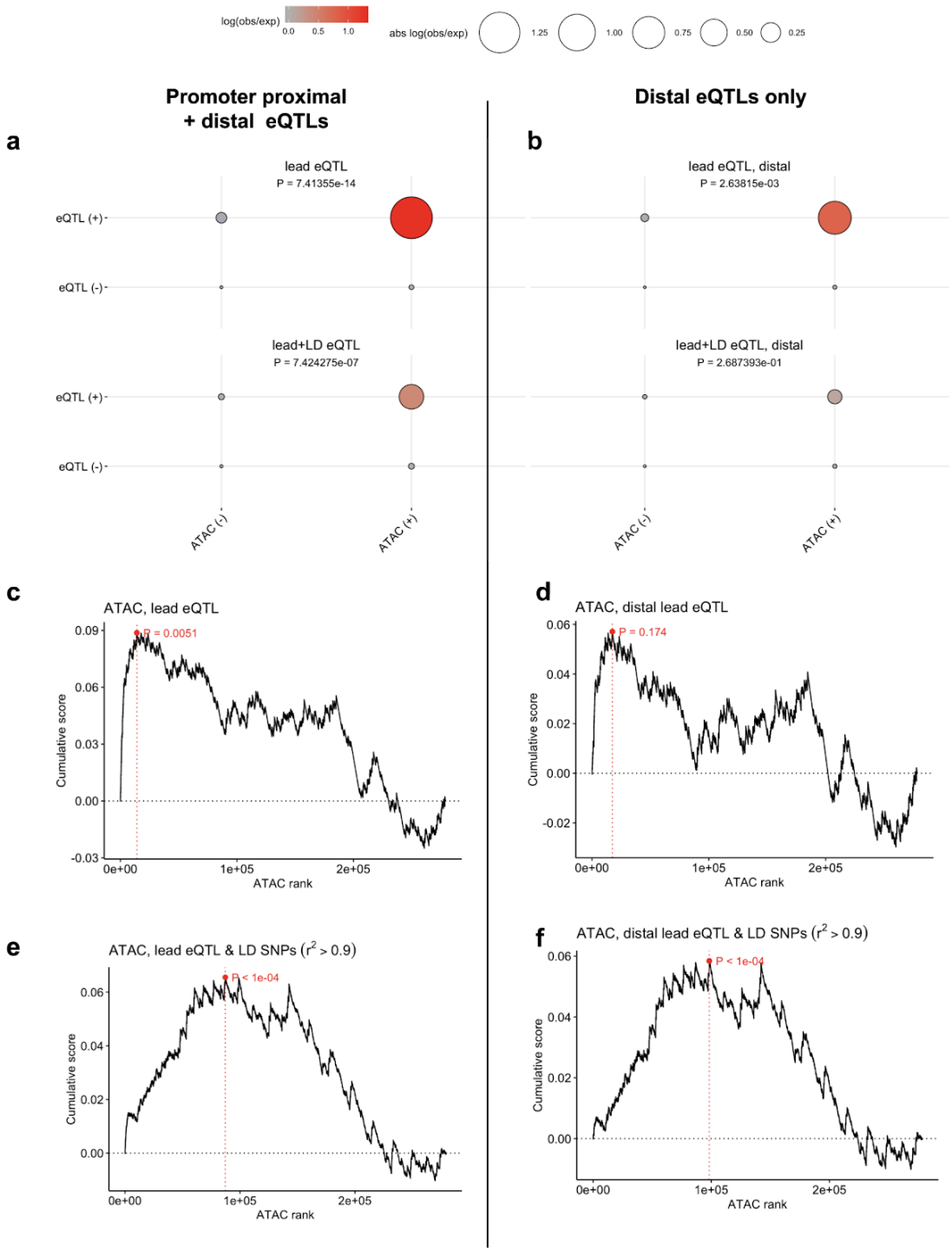


Figure 48: ATAC-seq features at eQTLs. Lead eQTLs and SNPs in strong LD ($r^2 > 0.9$) with lead eQTLs (LD SNPs) are considered. Overlap of ATAC-seq peaks with lead eQTLs (top) as well as overlap with lead eQTLs and LD SNPs (bottom) for all (**a**) and distal SNPs only (**b**). ATAC-seq signal enrichment in lead eQTLs for all (**c**) and distal SNPs only (**d**). ATAC-seq signal enrichment in lead eQTLs and LD SNPs for all (**e**) and distal SNPs only (**f**). p-value in (a,b) obtained by Chi-squared test. Distal SNPs are within +2,000 to -500 bp TSS distance. p-value in (c-f) obtained by permutation test. Maximum cumulative score, its rank and corresponding p-value in (c-f) indicated in red.

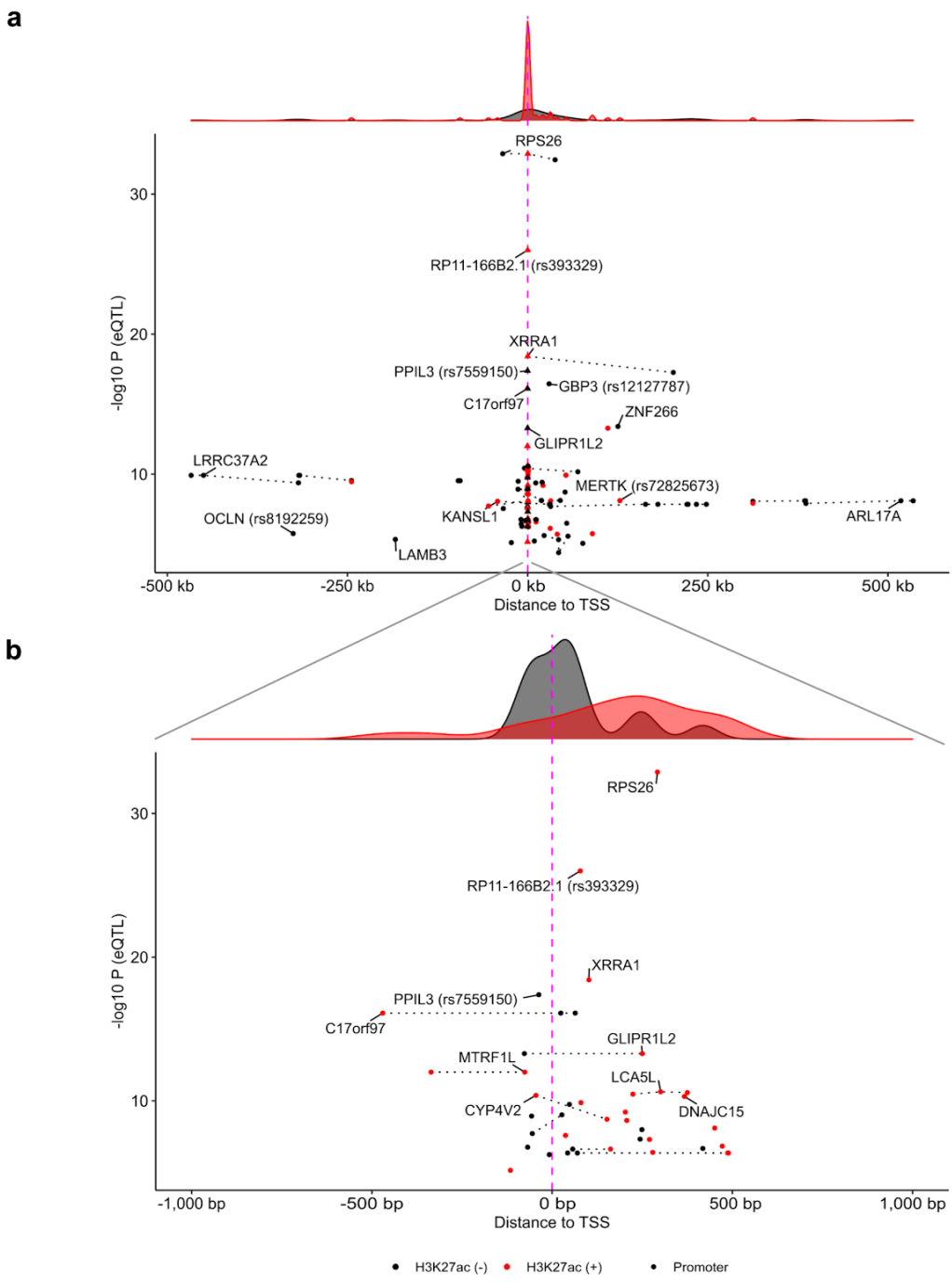


Figure 49: Priotitized candidiate cis-regulatory SNPs by distance to TSS of associated gene. (**a**) All candidate SNPs. SNPs with -log10(P) > 12 in eQTL mapping and those with absolute TSS distance > 100 kb are annotated by gene name. (**b**) Candidate SNPs in a 1 kb region around the corresponding gene's TSS. SNPs with -log10(P) > 10 in eQTL mapping are annotated by gene name. If multiple SNPs visible in the plotted region are prioritized for a given gene they are connected by a black dotted line. If a single SNP was prioritized, the SNP identifier is annotated as well. Density plots of SNP positions stratified by H3K27ac peak overlap are shown on top of (a) and (b).

5.2.3 eQTL survival analysis

To identify potential cis-regulation by germline variants that is linked to disease outcome we associated genotypes at lead eQTLs of the 163 identified cis-regulated genes with survival. A Cox proportional hazards regression model for overall survival time and status “deceased from disease” was used to estimate the effect of the lead eQTL genotype on patient survival. We encoded the SNP genotype by the number of alternative alleles and controlled the model for the covariates age, sex, MNA status, stage 4 status, tumor purity, tumor ploidy as well as cohort membership. A total of 163 tests were conducted and p-values were adjusted for multiple testing (Benjamini-Hochberg). Controlling the results at FDR 0.05 did not yield any significant association (Figure 50a). The genes ZP3, GNPDA2, FAM118A and THEM50B showed associations of nominal $P < 0.01$. Lead eQTL rs1799210 of the extracellular matrix component encoding gene ZP3 (Zona Pellucida Glycoprotein 3) yielded the smallest nominal p-value (0.0038). We repeated the Cox regression by encoding the three possible genotypes of rs1799210 as nominal variables and found increased hazard ratios for carriers of the alternative allele (C) relative to the reference alle (T) and the homozygous alternative genotype to show the strongest estimated hazard ratio (9.17) (Figure 50b). We seeked to identify eQTL genes that are differentially expressed between patients who deceased from the disease and patients without disease-associated mortality (Section 3.1.3) and intersected the two gene sets (Figure 50 c). In total we find 24 protein-coding genes to be differentially expressed and to also be cis-regulated by germline variation according to our genome-wide eQTL analysis. This set of genes comprises APC2, ATP13A4, C22orf43, CAPN9, CNKSR1, CYP4V2, DCXR, EBPL, EFCAB2, FAHD1, FAM86B1, GLIPR1L2, KIF6, LPPR1, LRRC28, LSG1, METTL21B, NSUN2, NUTM2B, PILRA, POLR2J2, SMTNL1, SRSF10 and ZNF429. Cox regression p-values of these genes ranged from P=0.079 (GLIPR1L2 / rs4533075) to P=0.99 (EBPL / rs112332160). None of the genes was found nominally significant at $P \leq 0.05$. Lead eQTL rs4533075 of GLI Pathogenesis Related 1-like 2 (GLIPR1L2) showed the smallest p-value among differentially expressed eQTL genes (P=0.079). We tested rs4533075 genotypes in a separate Cox-regression controlling for aforementioned model coefficients but did not find a nominally significant association (heterozygous [T/C] P = 0.423, homozygous alternative [C/C] P=0.077) (Figure 50e). Supplementary table 19 lists Cox proportional hazards regression results for the effect of lead eQTLs genotypes on overall survival.

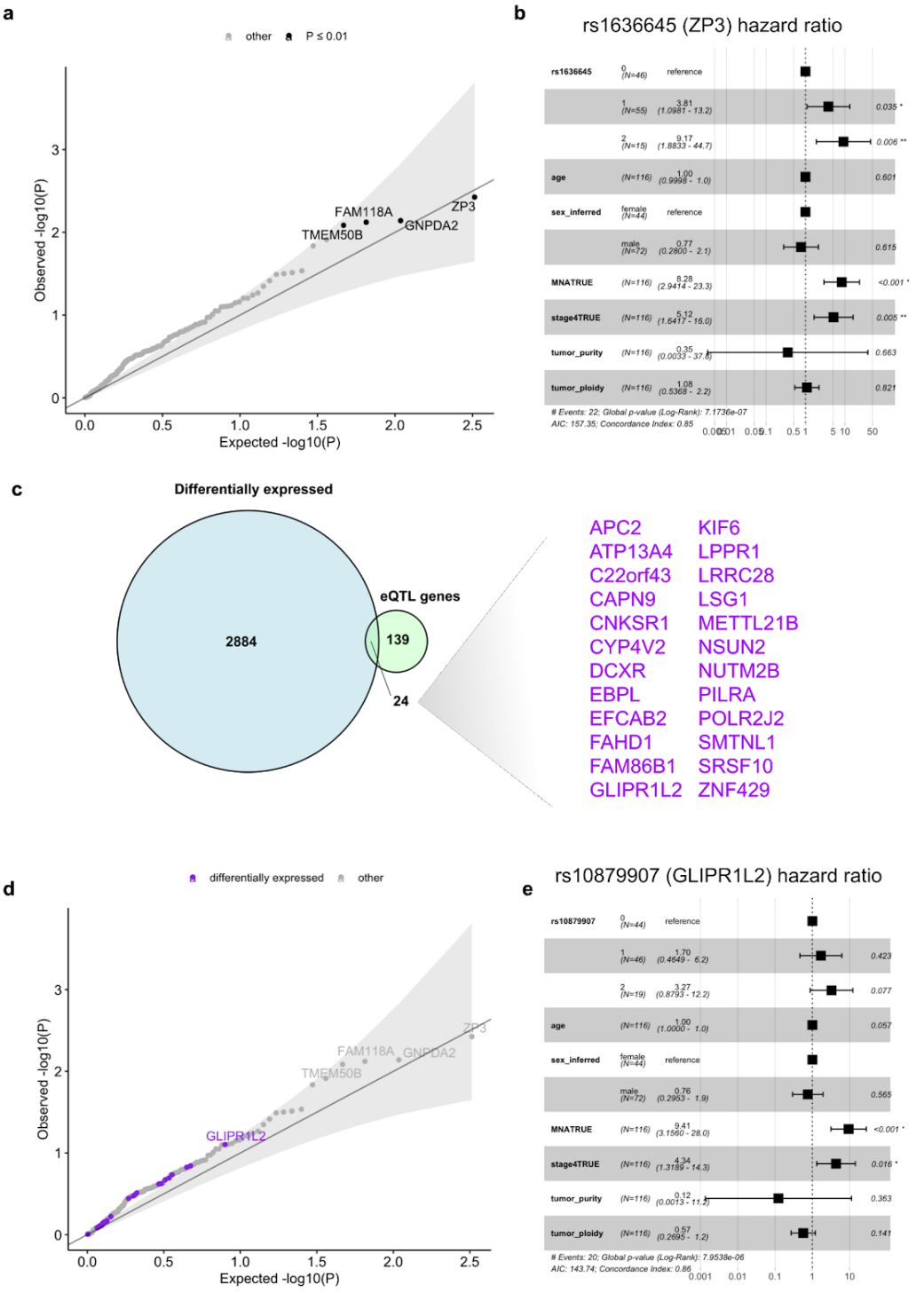


Figure 50: Association of lead eQTL genotype with patient survival. **a**, Observed and expected p-values of Cox proportional hazards regression. **b**, Estimated hazard ratios and p-values for model coefficients in the Cox model of ZP3 lead eQTL rs1636645. **c**, 24 eQTL genes are differentially expressed between patients who have died from the disease and others. **d**, Observed and expected p-values of lead eQTLs in differentially expressed genes. Top-ranking gene GLIPR1L2 is annotated. **e**, Estimated hazard ratios and p-values for model coefficients in the Cox regression of GLIPR1L2 lead eQTL rs10879907.

5.2.4 GWAS Quantitative trait loci at neuroblastoma genome-wide associations

To investigate cis-regulatory effects at neuroblastoma risk loci we analyzed our QTL mapping results in conjunction with GWA summary statistics from 2,101 neuroblastoma cases and 4,202 controls (McDaniel et al. 2017). The GWAS data is based on a limited set of variants analysed by SNP arrays and McDaniel and colleagues imputed genotypes to obtain associations for the extended set of 1000 genomes phase 3 SNPs. The genotypes we determine and use in QTL mapping in our 116 samples are based on the same set of SNPs (Section 3.1.4), and therefore we were able to integrate results of both analyses. Considering the thresholds MAF ≥ 1% and imputation quality score ≥ 0.7, as applied by McDaniel et al., the GWA statistics of 7,962,206 SNPs were available. The authors considered GWA results of $P < 5 \times 10^{-8}$ as significant hits and we applied the same threshold yielding 156 associations, to which we here referred as “GWAS risk SNPs”. These GWAS risk SNPs clustered at 6 distinct genomic loci (BARD1, CPZ, CASC15, HACE1/LIN28B, LMO1 and TP53), while signals at MLF1 and HSD17B12 were just below the defined discovery threshold (Figure 51a and MacDaniel et al. 2017 figure 1). Our QTL mapping is limited to SNPs with observed genotype differences between the 116 donors analyzed and to SNPs within cis-windows of expressed genes. We obtained eQTL statistics for at least one gene at 3,971,392 (49.9%) of GWAS SNPs and 87 (55.8%) of GWAS risk SNPs. To investigate overlap of GWAS associations at eQTL genes we determined a nominal eQTL p-value threshold defined as the weakest genome-wide lead eQTL association ($P = 5.15 \times 10^{-6}$). However, we did not find an overlap of risk SNPs in eQTL associations below this threshold. Similarly, we investigated associations below the corresponding aseQTL p-value threshold ($P = 1.49 \times 10^{-7}$). We identified rs2168101 as the only SNP significant in both GWA and aseQTL mapping. According to our analysis, the neuroblastoma risk SNP rs2168101 is a unique lead aseQTL association for LMO1. Our findining confirms the cis-regulatory effect of rs2168101, which was previously identified as a risk-associated intronic super-enhancer polymorphism altering a GATA family binding site (D. A. Oldridge et al. 2015). In concordance with Oldridge and colleagues we find rs2168101 to overlap an accessible chromatin region as defined by an ATAC-seq peak in SH-SY5Y. Figure 51b and c show GWA and aseQTL associations in a 50 kb window around the lead aseQTL rs2168101 as well as ATAC-seq and H3K27ac ChIP-seq signals at this genomic interval. Interestingly, we did not find an association between the genotype of rs2168101 and total expression of LMO1 (eQTL mapping P = 0.48) (Supplementary figure 12).

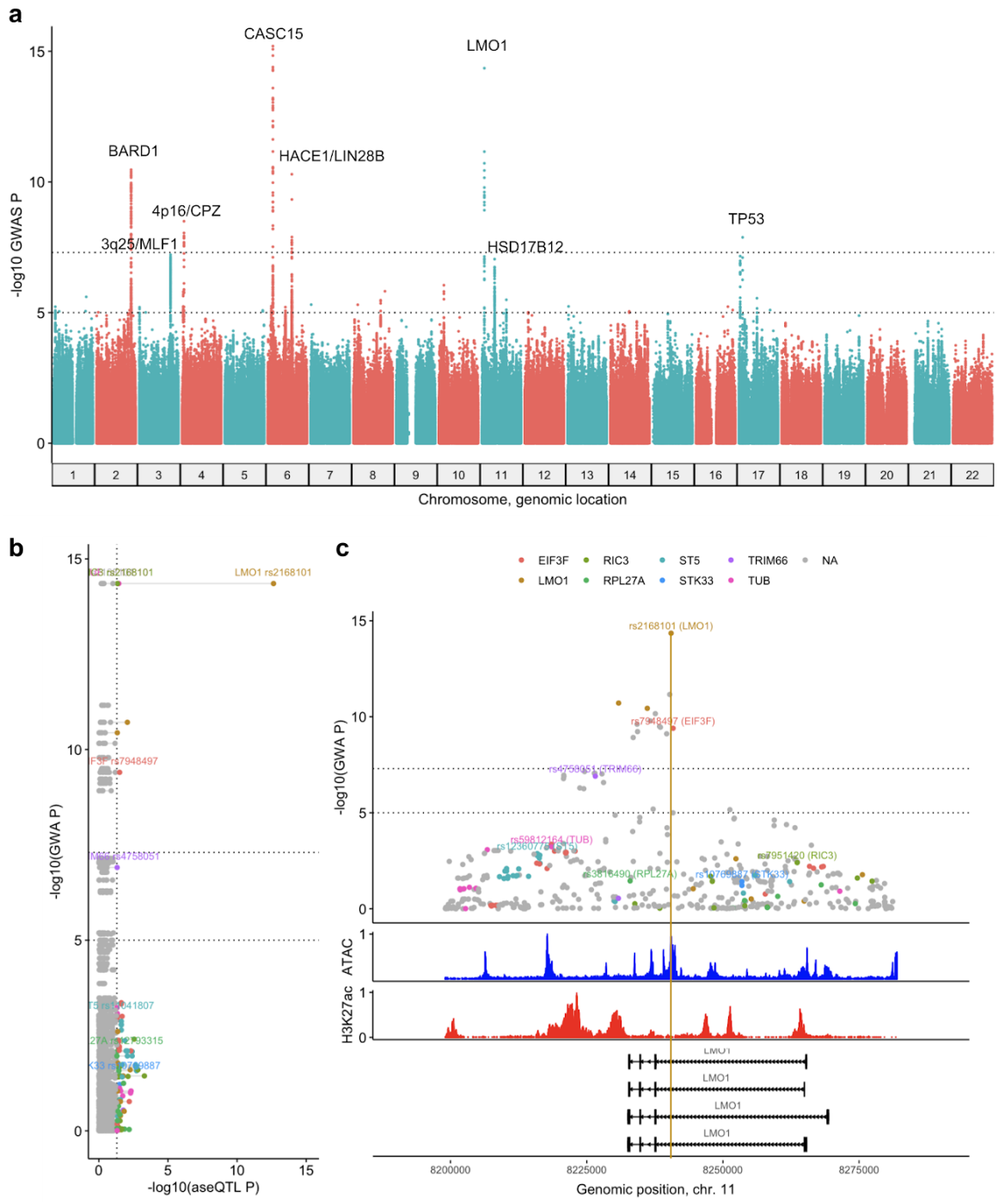


Figure 51: Overview of GWA p-values and aseQTL associations at the LMO1 locus. **a**, GWA between SNP genotypes and neuroblastoma susceptibility from McDaniel et al. 2017. p-value thresholds 5×10^{-8} and 1×10^{-5} indicated by dotted horizontal lines. **b**, GWA and aseQTL p-values for SNPs in a 50 kb window around the lead aseQTL rs2168101 at the LMO1 locus. aseQTL tests for the same SNP are connected by a solid grey line. aseQTL P = 0.05 (nominal) indicated by vertical dotted line. **c**, GWA p-value and genomic location of SNPs shown in (a); ATAC-seq and H3K27ac ChIP-seq signals by relative read coverage in neuroblastoma cell line SH-SY5Y. Overlap of ATAC peak and intronic location of rs2168101 indicated by yellow vertical line.

We further aimed to prioritize candidate genes that could be regulated by risk variants based on trends in eQTL statistics that did not reach genome-wide significance. As our sample size is small, elevated eQTL mapping p-values could indicate a regulatory effect, which we are unable to identify in a genome-wide context due to a lack of statistical power. We collected the strongest GWAS risk SNP for the BARD1, 3q25/MLF1, 4p16/CPZ, CASC15, HACE1/LIN28B, LMO1, HSD17B12 and TP53 locus from the summary statistic (Figure 51a). We then kept SNPs of GWAS $P < 1 \times 10^{-5}$ at a maximum distance of 1 Mb from the strongest associated SNP, resulting in a total of 884 SNPs across the eight loci. 556 of these SNPs had at least one eQTL test and a total of 3,070 eQTL tests were considered. We ranked these tests based on their nominal eQTL mapping p-value per risk locus. Figure 52 shows these eQTL ranks for the selected tests per risk locus. We prioritize the top ranking gene, for which we find nominally significant eQTL associations. Our analysis prioritizes cis-regulation of BARD1 at the BARD1 locus with multiple SNP tests below nominal eQTL $P < 0.05$ which clustered broadly around the BARD1 TSS. Figure 53 shows GWA and eQTL p-values at the BARD1 locus. Cis-regulation of LXN was prioritized at the 3q25/MLF1 risk locus, closely followed by GFM1 (Supplementary figure 14). HMX1 was prioritized at the 4p16/CPZ locus (Supplementary figure 15). Based on the eQTL alone we prioritize cis-regulation of RPL27A at the LMO1 locus. However, we did find total expression of RPL27A to be associated with the functionally characterized GWAS risk SNP rs2168101 (eQTL $P = 0.62$) (Supplementary figure 12). We prioritize cis-regulation of EXT2 at the HSD17B12 locus (see Supplementary figure 16). There were no nominally significant eQTL associations at the CASC15 and HACE1/LIN28B locus. And at the TP53 locus the GWAS SNPs considered were not part of any eQTL mapping and thus we could not prioritize any cis-regulatory effects. Intrigued by the marked agreement of LMO1 aseQTL mapping and its enhancer SNP, we used the same prioritization strategy based on aseQTL mappings at GWAS risk SNPs. Both methods agreed on the prioritization of BARD1 at the respective risk locus but there were differences in genes prioritized at the other risk loci (Supplementary figure17). However, above we showed that aseQTL mapping is affected by skews in genotype distributions (Section 5.2.1). These biases may influence the sensitivity to detect cis-regulatory effects between genes at a given risk locus. Thus we cannot reliably compare p-values at these loci, which makes it difficult to interpret prioritizations based on ranks of aseQTL p-values.

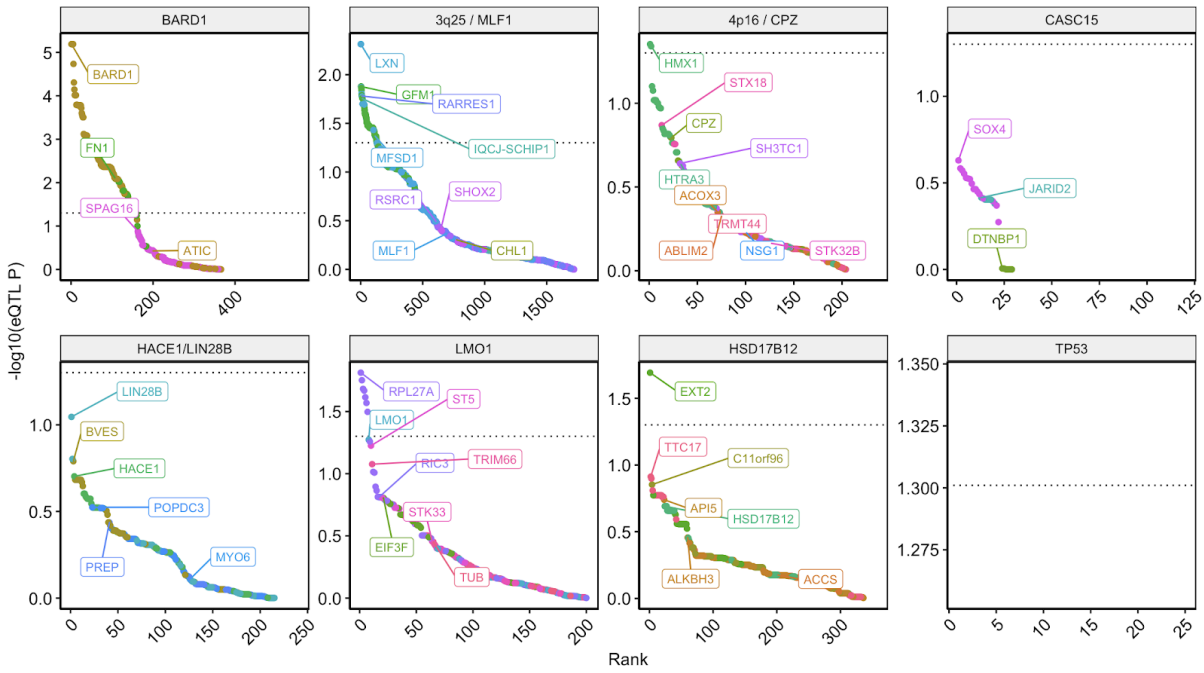


Figure 52: eQTL association tests at GWAS risk loci. Color indicates tested gene. Smallest eQTL association p-value per gene labeled by gene name. Dotted line indicates eQTL association p-value 0.05 (nominal). Only associations of SNPs informative for both GWA and eQTL analysis shown.

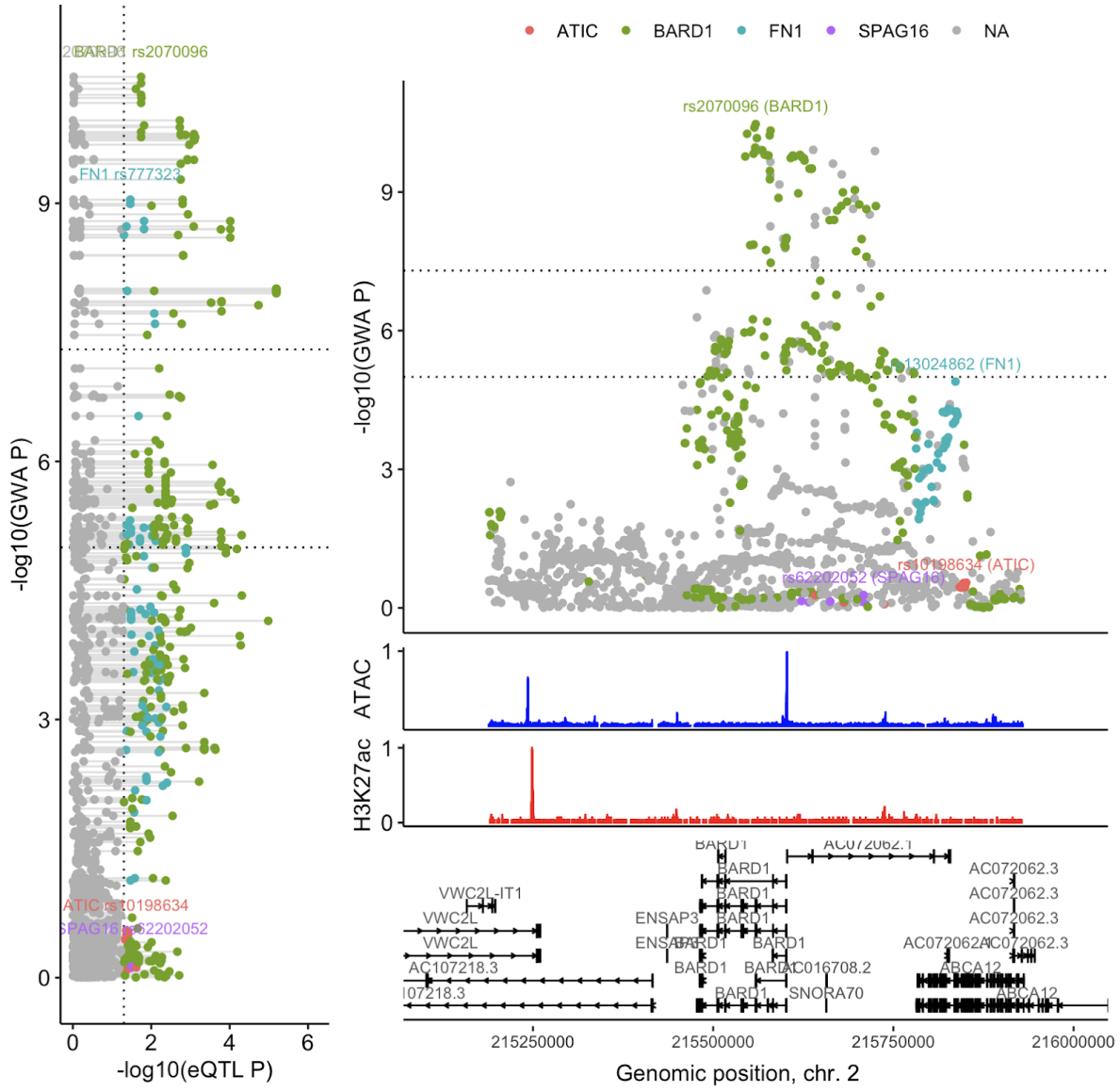


Figure 53: eQTL analysis at BARD1 risk locus. Left: GWA and eQTL association p-values. eQTL tests of different genes for the same SNP are connected by a grey line. Tests below nominal p-value threshold are color-coded by gene name, others in grey. Dotted line indicates threshold P = 0.05 (nominal). Right: Genome-wide risk association of SNPs from McDaniel et al. 2017. SNPs are annotated by gene with strongest eQTL association with P < 0.05 (nominal). SNPs without eQTL tests and those without tests P < 0.05 (nominal) in grey. H3K27ac-seq ChIP and ATAC-seq signals by read coverage from neuroblastoma cell line SH-SY5Y relative to maximum coverage at the locus.

5.3 Discussion

Cis-regulatory analysis by QTL mapping can provide useful information on the interplay between genes and germline genetic variation, such as SNPs involved in cis-regulation of gene expression. We mapped cis-QTLs of expression and allele-specific expression in 116 neuroblastoma tumors using a comprehensive set of WGS-based SNP genotypes and controlling for tumor-specific local somatic copy-number alterations among other covariates. We identified 163 and 24 cis-regulated genes in eQTL and aseQTL mapping respectively. We found QTLs frequently close to the transcription start site of associated genes inline with previous observations (Stranger et al. 2005; Veyrieras et al. 2008; Stranger et al. 2012), indicating that single promoter-proximal QTL variants typically have stronger influence on transcription than distal QTLs, such as those in enhancer elements. However, we cannot exclude that joint effects of multiple SNPs within such elements could have stronger effects, because our mapping strategy only considered the association between a single SNP and the quantitative trait per test.

Comparison between eQTL and aseQTL mapping

We found little overlap between the result from ASE- and total expression-based mapping strategies. Previous studies showed that genomic regions with allelic expression were enriched in cis-eQTLs in diploid cells (Ge et al. 2009), showing that genetic cis-regulation by germline variants leads to expression imbalances. However, different from our expectations we did not find aseQTL mapping of genes with sufficient ASE-informative samples to recapitulate eQTL associations, inline with previous comparisons of eQTL and aseQTL mapping in mouse adipose tissue (Hasin-Brumshtein et al. 2014). To identify causes for this discrepancy we analysed genotype distributions and found that subsetting to ASE-informative samples introduced deviations from the Hardy-Weinberg principle in genotypes considered in aseQTL association tests. Relative frequencies of heterozygous and homozygous genotypes at aseQTLs explained a substantial amount of the differences between p-values from aseQTL and eQTL mapping, indicating that dependencies between genotypes of aseQTL and ASE SNPs (instrument SNPs) may be the underlying cause of poor overlap between the two mapping strategies. This is in contrast to the conclusion drawn by Hasin-Brumshtein and colleagues, who attributed the differences mainly to trans effects influencing the eQTL mapping result. The authors investigated factors like genomic

background and sex specificity as possible alternative explanations, but to our knowlege did not determine the effect of genotype distributions in aseQTL mapping.

We propose that strong LD between the ASE SNP and the tested SNP impairs aseQTL mapping. This could e.g. occur if the heterozygous genotype of an ASE SNP (which is required to measure ASE in the first place), increases the likelihood of observing a heterozygous genotype in the tested SNP (Supplementary figure 11b). The resulting scarcity of homozygous genotypes at the tested SNP for ASE informative samples would then decrease the power of the association test, leading to false negative results. We concluded that eQTL mapping is not biased by this effect because the phenotype does not require subsetting to informative samples and therefore generally all samples are considered (Supplementary figure 11a).

Overlap of eQTLs with accessible chromatin and H3K27ac

We examined eQTLs in relation to epigenetic readouts of chromatin states associated with CREs (Section 2.2.1) in neuroblastoma cell line SH-SY5Y and found corresponding ATAC-seq as well as H3K27ac ChiP-seq peaks and signals enriched in the lead eQTLs identified. The enrichment was less pronounced when we included SNPs in LD ($r^2 > 0.9$) with lead eQTLs or considered promoter-distal SNPs only, suggesting that WGS-based genotyping and exhaustive eQTL association tests may already prioritize functional variation by association strength at a moderate sample size, especially if eQTL SNPs are proximal to the gene's TSS. However, as effects of distal SNPs might be weaker, our sample size is likely too small to reliably pinpoint many functional SNPs in distal CREs, such as enhancers. This could explain the weaker enrichment of ATAC-seq observations in distal compared to promoter proximal lead eQTLs. We cannot conclude that fewer or weaker associations of distal SNPs indicate weaker cis-regulatory effects of this form of variation, because multiple CREs tend to interact with a promoter to orchestrate tissue specific gene expression (Heinz et al. 2015). This implies that SNPs in several CRE can modulate expression at the same time and their combined effect could even be stronger than that of a single promoter-proximal variant. However, as our association test considers SNPs separately and LD between functional variants of enhancers in greater distance to each other is likely weak, combined effects are less likely not captured by our QTL mapping strategy. Another reason for weaker enrichment of ATAC-seq and H3K27ac ChIP-seq observations at distal lead eQTLs could be that our proxy (cell line SH-SY5Y) deviates from the actual epigenetic state of patient tumors. The epigenetic state of enhancer elements was found to be highly

tissue-specific (Blow et al. 2010; Visel et al. 2009; Ong and Corces 2011). eQTL mapping results may therefore vary even between cell types and this can even be observed in closely related tissues. For example, an eQTL study in melanocytes found that over a third of eQTL genes were not identified in a different eQTL study of skin tissue and those associations were linked to melanocyte-specific pathways and melanoma risk loci (T. Zhang et al. 2018). Consequently, if the cell identity differs between SH-SY5Y cells and patient tumors we would expect a weaker enrichment of enhancer-associated chromatin states at functional enhancer SNPs in patient tumors. If lead eQTLs correspond to true functional variation, this could explain the weaker enrichment we find for ATAC-seq and H3K27ac ChIP-seq observations at these SNPs. This problem can be addressed by integrating tumor-matched epigenetic assays into QTL mapping. Such an approach could also overcome an additional problem regarding epigenetic differences: Even if the cell identity between a proxy and tumor cells is the same, germline and somatic variation will very likely differ. Because genetic variation contributes to epigenetic heterogeneity (Section 2.2.3) we have to expect differences in chromatin states between cell lines and tumor samples. However, the major hurdle for integrating matched epigenetic readouts from tumors is that corresponding assays are not yet routinely applied in the molecular characterization of cancer samples of larger cohorts that are required e.g. for QTL mapping. Lastly, tissue-specific effects might be introduced by normal and immune cell admixture in tumor samples. Our QTL mapping is controlled for estimates of tumor purity which compensates for differences in the quantitative trait due to varying amounts of cancer cell fraction. However, if a gene is consistently and highly expressed in the normal and immune cell fraction, QTL mappings could reflect cis-regulation in the non-tumor-cells of the sample. As these cells will very likely have a cell identity different from tumor cells (immune cells, vasculature), functional lead eQTLs could lie outside of tumor-specific CREs.

Prioritization of candidate cis-regulatory SNPs

Regulatory SNPs interact with open chromatin regions and these variants also show enrichment in disease-associated loci (Maurano et al. 2012; Degner et al. 2012). We decided to prioritized candidate cis-regulatory variation in lead eQTLs and SNPs in strong LD ($r^2 > 0.9$) by their overlap with ATAC-seq peaks. From the 163 eQTL genes, 59 (36.2%) had at least one prioritized candidate cis-regulatory SNP (Supplementary table 18). Our method is based on the assumption that functional SNPs are preferentially located in tissue-specific accessible chromatin at CREs (Trynka et al. 2013; Farh et al. 2015; Onengut-Gumuscu et al. 2015; Lu Chen et al. 2016). The prioritized variants should

therefore reflect a set of higher confidence for a functional role of the SNP compared to a prioritization based on eQTL association alone, as the latter is additionally influenced by LD. For example, strong LD could lead to identical genotypes of two neighboring SNPs that show strong association with the quantitative trait of a gene in the cis-window. However, only one of the SNPs might be functional and the association of the other SNP is then solely based on LD with the functional SNP. If one of the two SNPs resides in an open chromatin region, this could indicate a modulation of a TF binding motif in a CRE. Hence, our method would prioritize the open chromatin variant. Using this approach the majority of our identified eQTL genes remain uncharacterized in terms of functional SNPs underlying the association. This could mean that in the majority of cis-regulated genes functional SNPs are located outside of open chromatin. However, considerations of discrepancies in epigenetic state between patient tumors and the proxy cell line used to identify ATAC-seq peaks that are mentioned above apply here too. Considering the strong enrichment of ATAC-seq peaks in lead eQTL, the lower enrichment among their LD SNPs and frequent promoter-proximal associations, we suggest to prioritize lead eQTLs in descending order by TSS distance in genes that lack ATAC-seq-based prioritization.

Our prioritization method for cis-regulatory variation does not select SNPs that are outside of ATAC-seq peaks in SH-SY5Y cells. To overcome the problem of missed prioritizations due to the selection of one specific cell line and its open chromatin regions, one could use a broader panel of cell lines instead. Ideally these cell lines should be derived from both MNA and non-MNA tumors and also reflect different cellular subtypes that were previously identified (Boeva et al. 2017). Yet, including multiple cell lines and subsequently more open chromatin regions could also inflate the group of false positive prioritizations. Additionally, genetic differences between tumors and cell lines may still lead to missing prioritization of functional SNPs. For example, let us consider an extreme case, in which a cis-regulatory SNP induces chromatin accessibility only for one of two possible alleles. If this allele is only found among tumor samples, but not in the selected cell lines, then the SNP cannot be prioritized.

To overcome these difficulties SNP-induced changes in epigenetic state can be predicted. Machine learning methods based on hidden markov models (Ernst and Kellis 2012), support vector machines (Fletez-Brant et al. 2013; Ghandi et al. 2014) and deep neural networks (Jian Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016; Quang and Xie 2019) were developed to predict epigenetic states based on sequence features. Such methods can

be used to predict differences in chromatin accessibility and transcription factor binding at sites of sequence variation. Fixed-length sequences (usually from the reference genome) and cell-type specific features of transcription factor binding (ChIP-seq) and chromatin accessibility (DNAse-seq, ATAC-seq) are used in the training step. Once trained, the algorithms predict feature scores for arbitrary sequences. Differences in scores may be used to prioritize sequences containing functional variants, such as SNPs that change the probability of transcription factor binding or chromatin accessibility. Thus, prediction-based prioritization is based on the observation that functional variants modulate the epigenetic state of CREs (Section 2.2.3). This resolves the extreme scenario described above, in which one allele is associated with closed chromatin and the other allele with open chromatin. If the predicted functional consequence is large enough the variant can be prioritized. Because in these methods many instances of chromatin features are learned genome-wide they are capable of prioritizing variants at alleles that may not even be present in the genome of the (epigenetic) training samples. However, the assumption that the reference genome sequence determines epigenetic features in those samples is an oversimplification that could lower prediction accuracy.

Missing prioritization of SNPs could also indicate that true functional variation was not considered in the analysis. Rare functional SNPs or germline structural and copy-number variants might be in LD with common SNPs at eQTL gene loci. We did not determine QTL associations of SNPs with MAF < 1% in a broader population (Section 3.1.4). Neither did we consider germline copy-number or structural variants. However, it was estimated that 40% or more of inheritable cis-regulation cannot be explained by common germline variation (Grundberg et al. 2012) and as much as 7% of eQTL associations could result from causal SVs (Chiang et al. 2017). To address these shortcomings of our analysis other variant types (e.g. germline SVs and CNVs) would have to be integrated. More permissive MAF thresholds or de-novo discovery of SNP could be used to determine rare variation from WGS. This way rare functional variants with large effect sizes could be discovered. But modest effect sizes would require larger sample sizes when results are controlled for the same FDR, because many more SNPs would need to be tested.

Association of eQTL genotypes and survival

We did not find significant associations of eQTL genotypes with disease-specific survival controlling for FDR 0.05 (Figure 50a), suggesting that major determinants of expression heritability in the identified eQTL genes are not strongly linked to survival. Smallest nominal

association p-values ($P < 0.01$) were found for lead eQTLs of genes ZP3, GNPDA2, FAM118A and TMEM50B. These genes do not have established roles in neuroblastoma or other malignancies. 24 differentially expressed genes overlapped with the set of eQTL genes. These genes include WNT pathway regulatory APC2, a paralog of the tumor suppressor APC, and human glioma pathogenesis-related protein like 2 (GLIPR1L2), a p53 target gene with a high degree of similarity to its homolog GLIPR1 (C. Ren et al. 2006). This paralog GLIPR1 is differentially expressed in variety of cancers (Murphy et al. 1995; Rich et al. 1996; C. Ren et al. 2004; Chilukamarri et al. 2007; Quinn et al. 2009; Awasthi et al. 2013) and a part of the TPX2-p53-GLIPR1 regulatory circuit, a modulator of key cancer hallmarks in bladder carcinoma (L. Yan et al. 2018). None of the lead SNPs in differentially expressed eQTL genes showed significant association with disease-specific survival, indicating that risk-related expression variability was mainly mediated by trans effects. We cannot exclude an aggregate risk effect of genetic regulation by eQTL SNPs, but would suggest to conduct QTL mapping in a larger cohort in order to identify survival-associated regulation.

eQTL and aseQTL mapping at the LMO1 enhancer SNP

In order to map regulatory effects at neuroblastoma susceptibility loci we examined eQTL and aseQTL mapping of risk loci identified in a published GWAS (McDaniel et al. 2017). We confirmed a strong aseQTL signal for the intronic enhancer SNP in LMO1 (D. A. Oldridge et al. 2015). This enhancer SNP (rs2168101) was not removed by the LD filter (Section 5.1.1) and subsequently selected as the representative SNP for gene-level association of LMO1. aseQTL association of rs2168101 was remarkably strong compared to surrounding SNPs and overlapped an SH-SY5Y ATAC-seq peak (Figure 51b,c). These findings confirm the regulatory role of rs2168101 for LMO1 and indicate that there is only a weak correlation of genotypes between rs2168101 and neighboring SNPs. Surprisingly, rs2168101 was not found to be an eQTL for LMO1 (Supplementary figure 12). Indeed, the association test did not even yield nominal significance, suggesting that this SNP is not a potent regulator of LMO1 steady state RNA levels in neuroblastoma tumors. We think that LMO1 regulation by genetic variability at this SNP is superimposed by trans regulatory effects but not other cis-effects, as ASE imbalance is still associated with rs2168101 heterozygousity. The question remains why rs2168101 confers susceptibility if it does not strongly regulate steady state RNA of LMO1 in tumors. We here try to give three possible explanations. First, the cis-regulatory effect of rs2168101 on LMO1 total RNA could be specific to a developmental state: While a trans effect on LMO1 is dominant in the tumor tissue, the SNP-mediated cis-effect might be a driver of total LMO1 expression in cells of the developmental nervous

system at the time of malignant transformation. eQTL can have opposing gene expression effects even between closely related tissues (Mizuno and Okada 2019), suggesting that the cell's regulatory program can modulate eQTLs effects. Similarly, somatic alterations that are acquired after the transformation may change the regulatory program such that a trans effect dominates control of LMO1 in tumors at the time of diagnosis. In fact it is already established that somatic alterations, such as MYCN amplification and 17q gain are key drivers of the regulatory program in neuroblastoma cells (Decaesteker et al. 2018; Zeid et al. 2018) and these changes could occur after rs2168101 has conferred its effect. Second, rs2168101 could modulate a temporal pattern of LMO1 expression, that is perhaps linked to the cell cycle. If this change is critical for disease initiation or progression the SNP may be effective in the regulatory context of a diagnostic tumor. However, the net effect of this time-dependent regulation must still be small enough to remain hidden in steady state gene expression as measured by batch RNA-seq, because otherwise we could have identified rs2168101 as an eQTL for LMO1. Third, rs2168101 could confer its effect by other means than regulation of LMO1. However, we do not detect any nominal significant eQTL gene association of rs2168101 (Supplementary figure 12). Yet, the SNP could affect expression of genes outside of the cis-window used in the eQTL mapping (Section 5.1.1), or it may be relevant to a disease trait in cis or trans that we do not consider. In T-cell acute lymphoblastic leukaemia cells LMO1 was found to be upregulated by somatic functional enhancer mutations (Z. Li et al. 2017), suggesting that an increase of LMO1 total RNA levels may indeed confer a growth advantage in some cancer cells. However, the reason for the missing association between the enhancer SNP and LMO1 total expression level in neuroblastoma tumors remains to be investigated.

Prioritizing genes at GWAS SNPs

We prioritized cis-regulation at eight GWAS risk loci based on the eQTL association p-values. None of the associations showed genome-wide significance, but we found nominally significant associations at the BARD1, 3q25/MLF1, 4p16/CPZ, LMO1 and HSD17B12 risk loci. A group of strong GWAS SNPs that broadly clustered around the BARD1 TSS showed nominal significant association with BARD1 expression (Figure 53). Earlier studies found risk SNPs at this locus to modulate expression of the oncogenic isoform BARD1β in which two exons are skipped (Bosse et al. 2012). But risk SNPs were also associated with differences in expression of the BARD1 full length transcript in several cell lines (Capasso et al. 2013). Our results provide additional evidence for risk-SNP-mediated cis-regulation of BARD1 by characterizing this effects in neuroblastoma

primary tumors. We did not consider isoform- or exon-level expression and thus SNP effects on these traits were not investigated. Our eQTL analysis prioritizes cis-regulation of LXN and GFM1 at the 3q25/MLF1 locus, HMX1 at the 4p16/CPZ locus, RPL27A at the LMO1 locus and EXT2 at the HSD17B12 locus (Figure 52). LXN is a homolog of retinoic acid receptor responder 1 (RARRES1), which is located just 24 kb upstream of LXN. Both LXN and RARRES1 were found to be coordinately downregulated in prostate cancer cell lines, and their repression was associated with increased invasiveness in primary epithelial prostatic cell cultures (E. E. Oldridge et al. 2013). HMX1 is a transcription factor that regulates the noradrenergic sympathetic cell fate (Furlan et al. 2013), indicating that cis-regulation at this risk locus might predispose to neuroblastoma through impairment of lineage decisions in the developing sympathetic nervous system. EXT2, which was prioritized at the HSD17B12 locus was previously described as a potential tumor suppressor in osteochondromas (Hecht et al. 1997; Philippe et al. 1997; Wuyts et al. 1998; Park et al. 1999; X.-J. Chen et al. 2016). However, we found that the strongest GWAS SNPs did not show nominally significant eQTL associations for HMX1, RPL27A and EXT2 at their respective risk loci (Supplementary figure 12, 15 and 16), making interpretation of these results challenging. These findings indicate that cis-regulation of HMX1 and EXT2 may not be the primary cause underlying the GWA, but a weaker regulatory signal undetected at stronger GWAS risk SNPs could cause the association. Generally, this could also be true for successful cis-eQTL prioritizations at the strongest risk SNPs. It was found that strong expression cis-heritability showed smaller effects on complex traits (D. W. Yao et al. 2020) and that human disease genes are depleted in cis-eQTLs, because they harbored larger and more robust regulatory domains (Xinchen Wang and Goldstein 2020). These findings suggest that cis-eQTL effects that underlie trait associations can be much smaller than effects on secondary genes (that do not underlie the GWA) and that extensive CREs are a mechanism of “eQTL robustness”. Alternatively, missing eQTL associations at top eQTL SNPs could also be explained by the hypothesis of differential regulatory programs between disease-initiating cells and the tissue in which eQTLs were mapped, as discussed for the LMO1 locus above.

When we prioritized cis-regulation at risk loci by aseQTL- instead of eQTL mapping we recapitulated prioritization of BARD1, but this approach selected different genes for most other risk loci (Supplementary figure 17). Interestingly, aseQTL-based prioritizations often selected genes closer to the risk SNPs compared to the eQTL-based method. This was the case for LMO1 (as discussed above) but also LIN28B and HSD17B2 at their respective loci. However, because of the biases we identified in aseQTL mapping (Section 5.1.1), we cannot

reliably compare its association strengths between genes, which hampers the utility of this approach. Understanding the exact mechanism by which risk loci confer neuroblastoma susceptibility requires further investigation. Larger sample sizes and methods that allow to compare relative association strengths in aseQTL mapping could help to uncover regulatory effects of disease associations. Single cell assays in developing neuronal cells may shed light on the interplay between genetic variants and gene expression. These methods could help us to investigate variant effects in a regulatory context that more closely resembles the one at the time of disease initiation.

Summary

In summary, we here provided a comprehensive map of genes subject to cis-regulation by germline variants in neuroblastoma tumors. Comparison of eQTL and aseQTL mapping results uncovered differences that were in large explained by deviations from the Hardy-Weinberg-principle due to subsampling to ASE-informative samples. We therefore decided to base our further investigations on eQTL mapping, where we found marked cis-effects close to TSSs and an enrichment of chromatin accessibility and H3K27ac for the strongest eQTL associations. Overlap of eQTL SNPs with ATAC-seq peaks provided a list of candidate functional cis-regulatory variants, that can guide the study of CREs and TFs involved in the regulation of the genes we have identified and cis-regulation in neuroblastoma in general. We did not detect risk-associated effects for individual lead eQTL controlling for FDR, but identified a list of differentially expressed genes that are subject to cis-regulation in primary tumors, that lead to promising candidates and pointed towards disease mechanisms subject to expression heretibility in cis. Our integrative analysis of GWAS and cis-QTL results confirmed the regulatory potential of the risk-associated LMO1 enhancer polymorphism rs2168101 and priotitized gene candidates for further investigations of cis-regulatory effects linked to neuroblastoma risk variants.

6 Conclusion and perspectives

In this work genetic and cis-regulatory effects in gene regulation in the childhood cancer neuroblastoma were investigated. Germline and somatic variation in 116 primary tumors were characterized and associated with total and allelic differences in gene expression. The analyses identified genetic and cis-regulatory effects associated with survival and disease pathways. CN dosage effects were found in telomere maintenance and other cancer pathways as well as in survival-associated allelic dosage effects in imprinting. My work highlights dosage-dependent regulation by SCNAs as a key mechanism of genetic deregulation in neuroblastoma that dominates other local genetic effects. SCNAs were found to regulate expression through dosage effects of chromosomal- and extrachromosomal DNA. Dosage effects of circular DNA were exclusive to large megabase-sized ecDNAs associated with strong CN amplifications. The dosage effects were absent for smaller but highly abundant kilobase-size eccDNAs. My work highlights the role of large ecDNAs in dosage-dependent upregulation by mono-allelic amplifications and shows that this form of genetic regulation controls expression of genes beyond MYCN and its co-amplified chromosomal neighborhood. Through cis-aseQTL mapping I confirmed cis-regulation by a previously identified LMO1 enhancer polymorphism. The integrative eQTL mapping of germline regulation and epigenetic readouts provides a valuable resource for the prioritization of functional non-coding variants in neuroblastoma gene regulation.

The sparsity of somatic coding mutations and the strong regulatory effect of SCNAs that were observed in neuroblastoma tumors point towards CN dosage effects as a key driver of this disease in both genomically more stable MNA tumors (by targeted amplification of MYCN on large ecDNAs and also 1p loss) and in many genomically unstable tumors by a variety of abundant segmental SCNAs across the entire genome, including frequent losses of 11q and strong 17q gains in non-MNA high risk tumors. Thanks to a limited set of affected genes the identification of focal amplification targets is relatively straight forward. Conversely, larger numerical and segmental SCNAs can comprise thousands of genes, which complicates identification of individual oncogenes and tumor suppressor genes deregulated by gains and losses respectively. I showed that quantification of dosage effects helps to identify genes and pathways deregulated by large SCNAs. For example, the results indicate that 17p loss causes downregulation of neuronal pathways in tumors of deceased patients by CN dosage effects. Dosage effects can propagate in signals to regulate genes in

trans, as I showed for 11q loss-mediated upregulation of distal histone genes H3F3B and H2AFJ in ALT tumors. To better understand trans-regulatory mechanisms we will need to describe individual trans-acting factors that are deregulated in cis. Future work should therefore focus on integrating dosage effects with regulatory networks to identify CN dosage sensitive factors on critical chromosome arms, such as 11q. More generally, the integration of CN dosage effects with cancer-specific regulatory networks may help to pinpoint individual driver genes deregulated by large SCNAs in genomically unstable tumors across many cancer types. In the future, such studies may lead to new therapeutic interventions in tumors driven by larger SCNAs that lack somatic mutations or amplifications associated with specific cancer genes. Computationally inferred candidate regulatory factors can be validated by shRNA- or CRISPR-based protocols. CRISPR screens have the advantage of a higher throughput as they can target many loci simultaneously. Indeed, CRISPR screens have already been used to characterize gene-dependencies of cellular phenotypes in neuroblastoma (Liying Chen et al. 2018; Durbin et al. 2018). In the future this technology will help to identify regulators of important but yet not well characterized molecular phenotypes in cancer, such as e.g. ALT in neuroblastoma.

ASE analysis is an important tool to discriminate local and cis- from trans-regulatory effects. The strength of ASE lies within its ability to identify genetic and cis-regulation even if it is superimposed by strong trans effects. For example, ASCN effects explained more of the variance in ASE than effects of total CN explained in the variance of total expression (Figure 21), likely because ASE directly captures the allelic expression differences resulting from copy-number dosage effects in somatic deregulation and is not affected by potentially compensatory regulation in trans. Similarly, I proposed that the strong cis-aseQTL association but the missing cis-eQTL association of LMO1 and its enhancer SNP rs2168101 could be due to extensive trans-regulation of this gene. Furthermore, ASE is able to uncover cis-regulation independent from its source. By correlating ASE with total RNA levels I could show that the imprinted gene RTL1 is regulated by expression differences between the two alleles. Based on previous reports of imprinting at this locus and the lack of genetic effects that could explain the observed expression differences I suggested that loss of imprinting underlies this effect. However, methylation at the RTL1 locus was not directly investigated. This example demonstrates how ASE helps to discover cis-effects by even uncharacterized sources of regulation. Still, such findings then require further validation. Thus, I suggest that future investigations should examine allele-specific methylation in neuroblastoma tumors,

specifically examine the broader DLK1-RTL1-DIO3 locus and associate its allele-specific methylation with expression differences and patient survival.

A disadvantage of ASE analysis is the varying statistical power between genes due to different numbers of samples that are informative for this phenotype. This makes it particularly difficult to compare gene-level statistics between ASE-based associations and it can hamper genome-wide testing in relatively small cohorts as ours, as perhaps reflected by the low number of cis-aseQTL genes that were identified. Therefore, under such circumstances it seems more appropriate to first perform genome-wide associations of total gene expression and then subsequently examine ASE effects to characterize cis- and genetic regulation in the identified genes. However, compared to ASE the discovery of cis- and genetic effects based on total gene expression may show reduced sensitivity in the detection of cis-regulation, particularly in genes that are additionally affected by strong regulation in trans (e.g. LMO1 locus). Furthermore, trans-regulation could mimic regulation in cis, in cases where the trans-regulatory factor is under control of the same genetic or cis-effect to be associated with the target locus. For example, the dosage effect analysis presented in chapter 3 is based on association between total gene expression and SCNA. However, if a trans- regulatory factor resides on the same CN segment this could lead to associations that are due to differences in expression of the trans-factor on the same segment. In such cases, ASE is a valuable tool to discriminate between cis- and trans-effects, as I showed for genes regulated by 11q-loss that were either regulated by local dosage or trans-regulatory effects (Figure 30a). Another disadvantage of the ASE phenotype was evident when considered in cis-QTL mapping: Due to correlations between genotypes of instrument and candidate effect SNP, cis-aseQTL testing introduces biases in genotype distributions. I suggest that such biases underlie missing cis-aseQTL associations of cis-eQTLs, especially in genes for which there exists a sufficient number ASE-informative samples. With growing sample sizes such biases could become negligible, but this issue makes aseQTL mapping less applicable to cohorts with limited sample availability, which is expected in rare diseases, such as specific childhood cancers. Furthermore, there is evidence of clonally fixed and dynamic monoallelic expression in individual cells (Reinius and Sandberg 2015) and its effect on RNA-seq in tumor tissue is not well understood. Dynamic mono-allelic expression, in which single cells exclusively or predominantly express a given gene from one of the alleles only, is likely averaged-out in bulk sequencing of tissue samples. However, if these effects become clonally fixed and the clone expands in the tumor cell population, then these effects are expected to impact ASE readouts in the bulk sample.

In such a scenario a portion of the variance in ASE that is not explained by genetic or epigenetic factors might result from clonally fixed random allelic expression. Further investigations are required to describe effects of random allelic expression in cancer cells. Such studies could for example compare ASE in single cell RNA-seq to ASE from bulk RNA-seq or pseudo-bulk from the same single cell experiment. However, they would require relatively high coverage in single cell RNA-seq for individual genes.

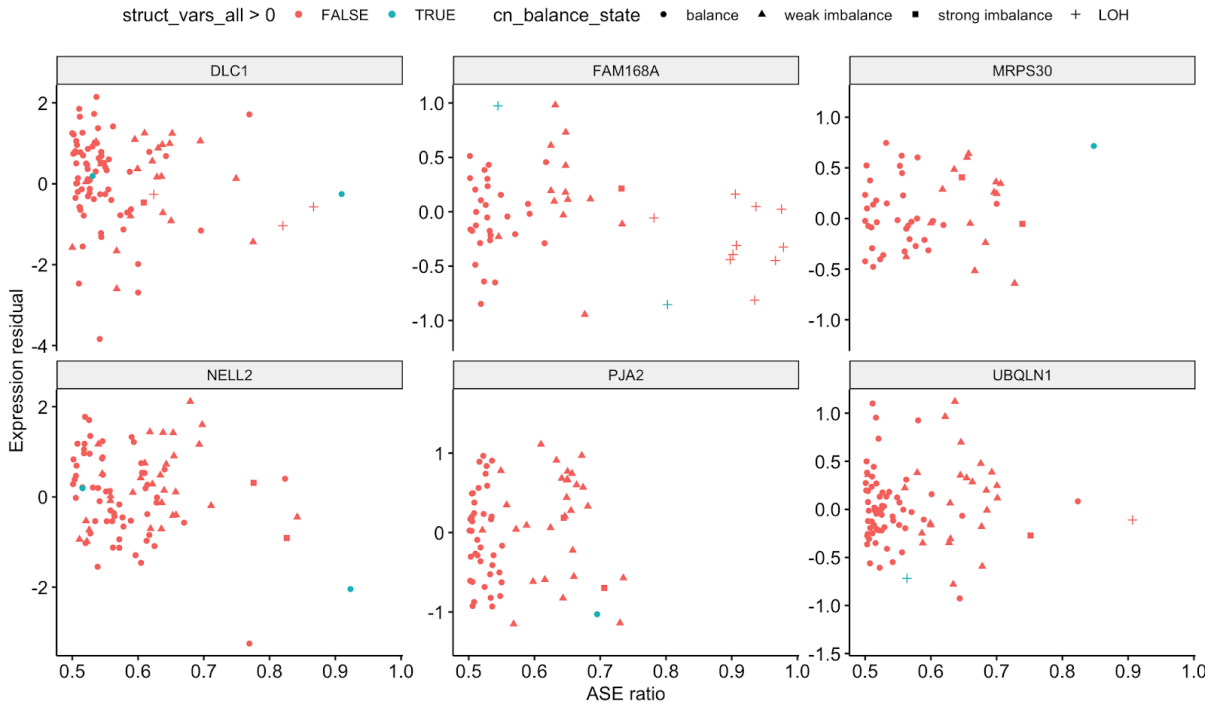
This work is limited by its focus on protein coding genes, a perhaps conservative definition of cis-regulation, and in that it does not address clonal heterogeneity and tumor evolution. Expression and ASE and consequently all associations with these phenotypes were only considered for protein coding genes. However, non-coding RNAs may have important roles in neuroblastoma biology (Molenaar, Domingo-Fernández, et al. 2012; Pandey et al. 2014; Russell et al. 2015). Thus, with appropriate sequencing protocols also regulation of non-coding RNAs should be investigated in future studies. Infact, a subset of the samples considered in this work (the Terminate-NB cohort) was profiled by an RNA-seq protocol that is not limited to mRNAs and therefore allows to quantify a variety of non-coding RNAs. However, I decided to only consider protein coding genes, because the other part of the samples analyzed (Peifer et al. 2015) was profiled by mRNA sequencing, which specifically enriches for this class of genes. Furthermore, the prioritization of functional variation in chapter 5 is based on established epigenetic characteristics of CREs. However, non-coding RNAs may also participate in regulatory mechanisms in cis (Gil and Ulitsky 2020) and if variants of such transcripts modulate their cis-regulatory potential, certain non-coding regulatory variants can be missed by focusing on CRE epigenetic marks only.

Intra-tumor heterogeneity (ITH) underlies darwinian selection of subclones in tumors (Gerlinger et al. 2012) and evolutionary trajectories in childhood cancers (Karlsson et al. 2018). In neuroblastoma, cellular heterogeneity may also arise from plasticity in regulatory states of individual cells (Boeva et al. 2017). Typically, ITH can be examined by sampling as a tumor in different regions and evolutionary processes may be studied by longitudinal sampling, as previously shown e.g. for diagnosis and relapse timepoints in neuroblastoma (Schramm et al. 2015). However, in this work regional or longitudinal tumor samples were not investigated and ITH was not examined specifically. Still, differences in genetic features of subclones within the tumor sample could potentially confound identification of somatic variation in our analysis. For example, SCNAs that are found in a subset of cancer cells only would result in an underestimation of the strength of a CN loss or gain in the bulk sample. Or

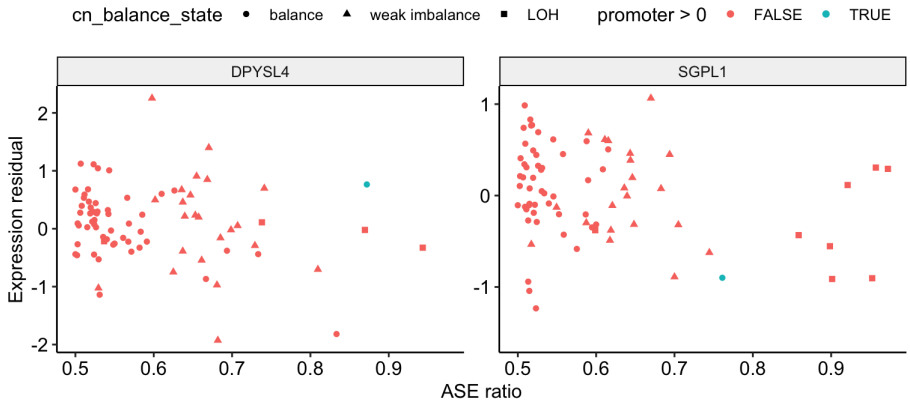
the alteration may even remain undetected if the subclonal population is too small. This could become especially problematic when longitudinal traits, such as survival endpoints, are considered, because a subclone harboring an undetected driver variant may expand in the course of the disease and could be causally related to the observed outcome. Similarly, driver variants could arise after the analyzed biopsy was collected. I referred to this concept in section 3.3, where late TP53 mutations as a second hit in 17p LOH samples were discussed as a possible cause for the high rate of mortality that is observed among the donors of tumors with this aberration.

My work provided a comprehensive analysis of the genetic and cis-regulation of gene expression in neuroblastomas and mechanistic links between gene regulation and quantitative and complex disease phenotypes. The integration of tumor-derived epigenetic profiles with regional and longitudinal data from tumor samples will help us in the future to understand the exact regulatory context of the cell of origin underlying disease initiation in individual tumors and the transitions of regulatory programs in the course of the disease. If possible, continued monitoring of somatic events and their regulatory consequences will guide treatment decisions aimed at improving survival. This way, we will hopefully soon be able to cure many patients with high-risk neuroblastoma with targeted, personalized treatment plans.

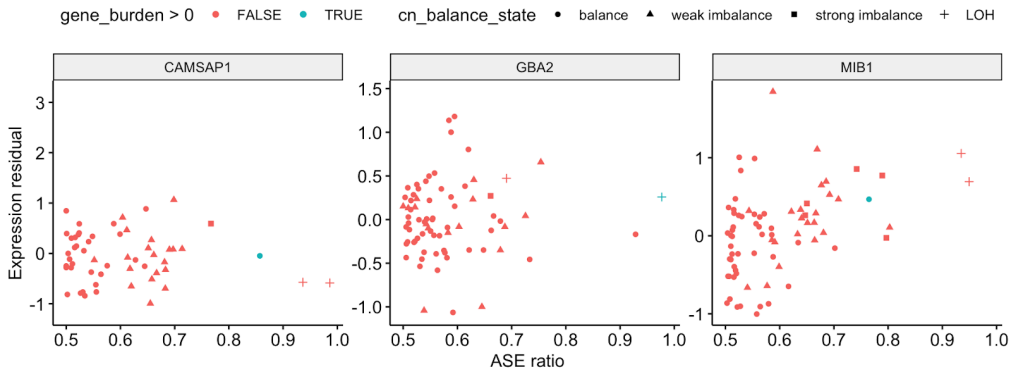
Appendix A: Supplementary figures



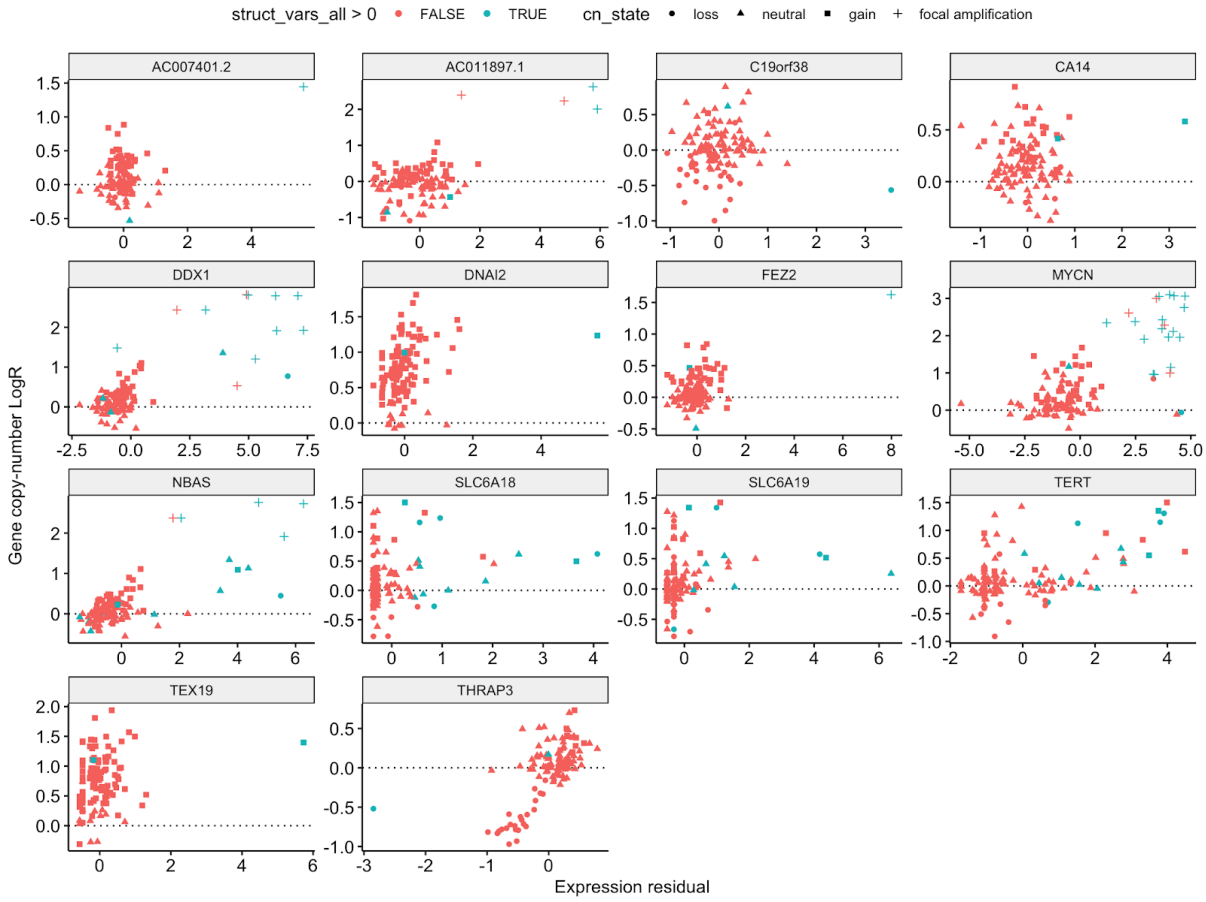
Supplementary figure 1: Expression and ASE ratio of genes with significant structural variation ASE variance component at (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with structural variants at gene coordinates +/- 100 kb flanking region.



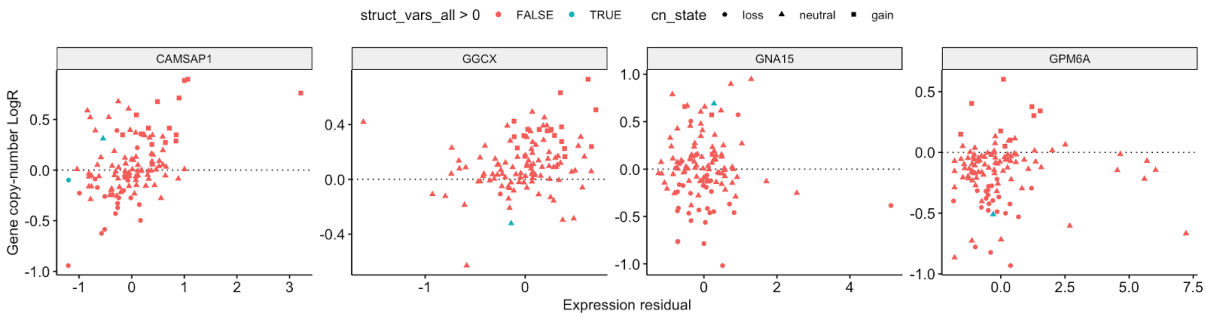
Supplementary figure 2: Expression and ASE ratio of genes with significant promoter SNV ASE variance component (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with SNV variants at the gene promoter.



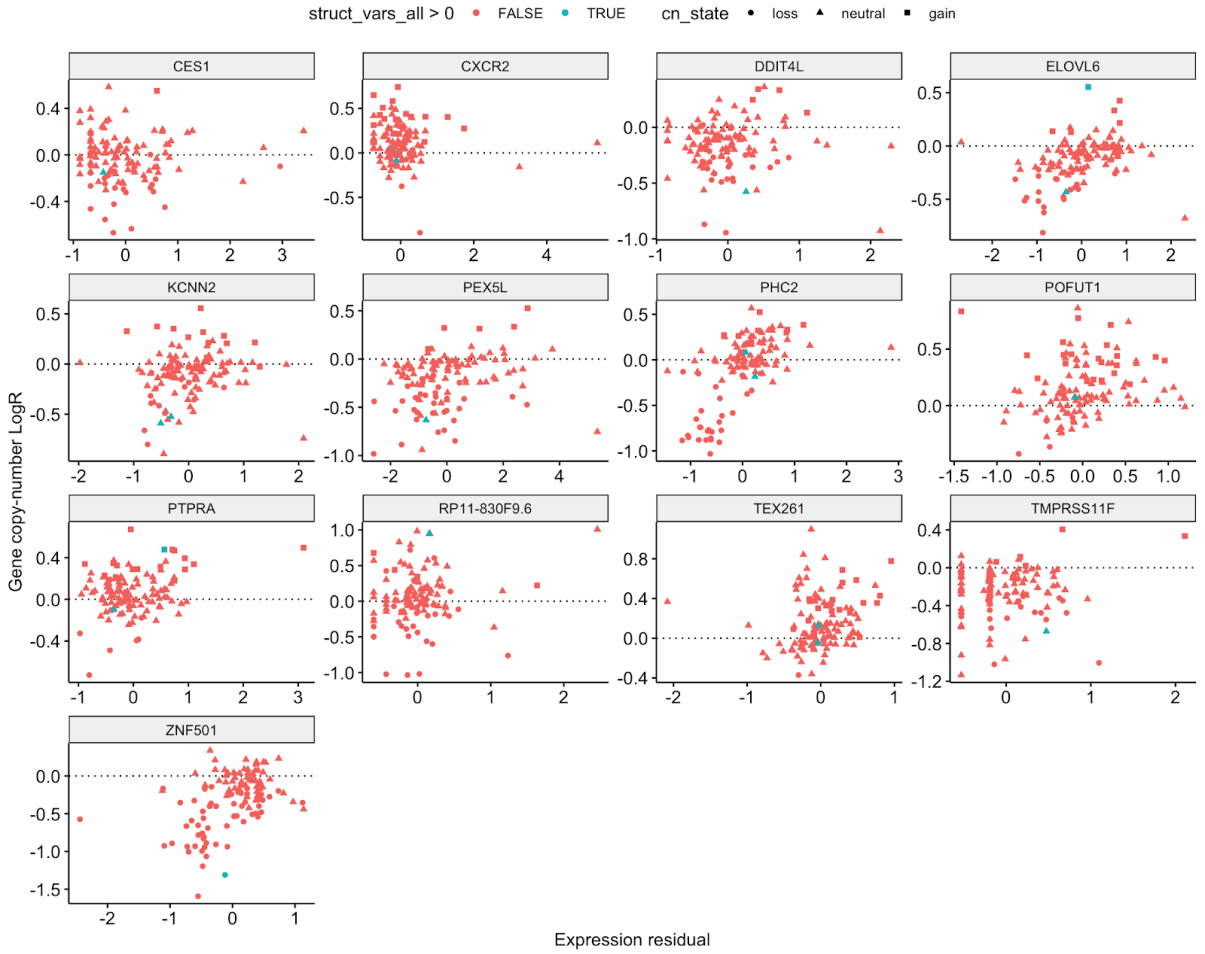
Supplementary figure 3: Expression and ASE ratio of genes with significant gene somatic SNV ASE variance component (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with somatic SNV variants at gene coordinates.



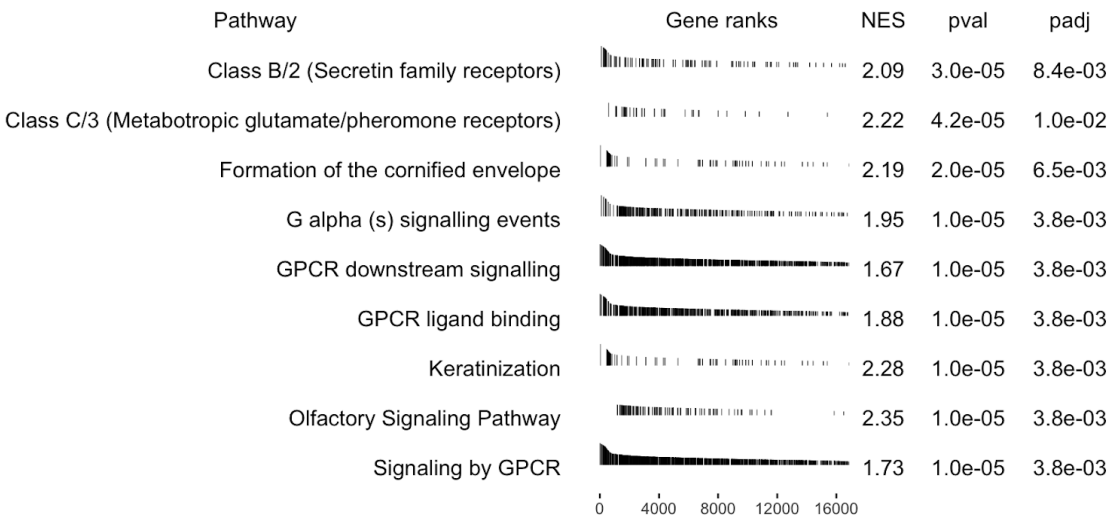
Supplementary figure 4: Expression and LogR of genes with significant structural variation expression variance component (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with structural variants at gene coordinates +/- 100 kb flanking region.



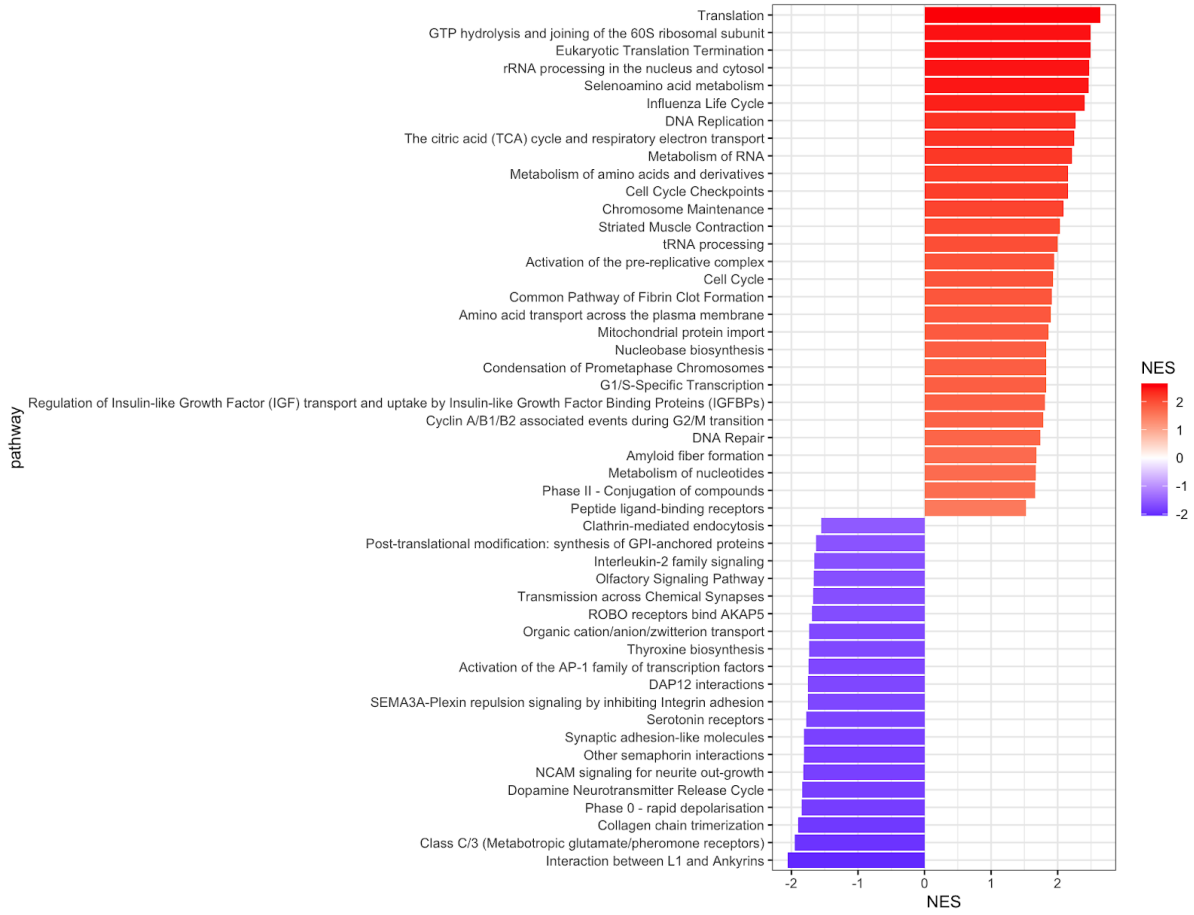
Supplementary figure 5: Expression and LogR of genes with significant promoter somatic SNV expression variance component (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with somatic SNV variants at the gene promoter.



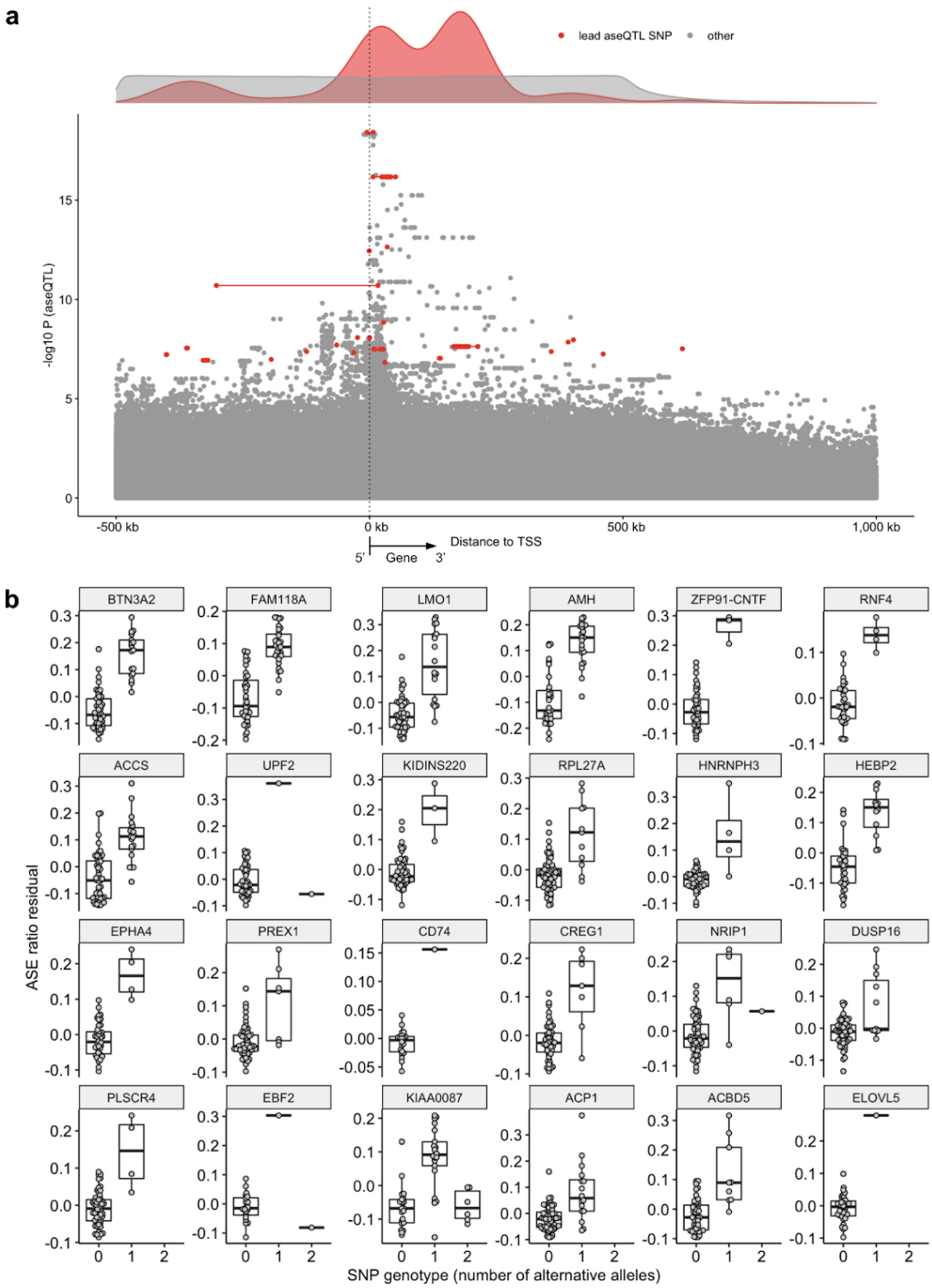
Supplementary figure 6: Expression and LogR of genes with significant gene somatic SNV expression variance component (FDR 5%, Benjamini Hochberg). Each dot represents a sample. Turquoise dots indicate samples with somatic SNV variants at gene coordinates.



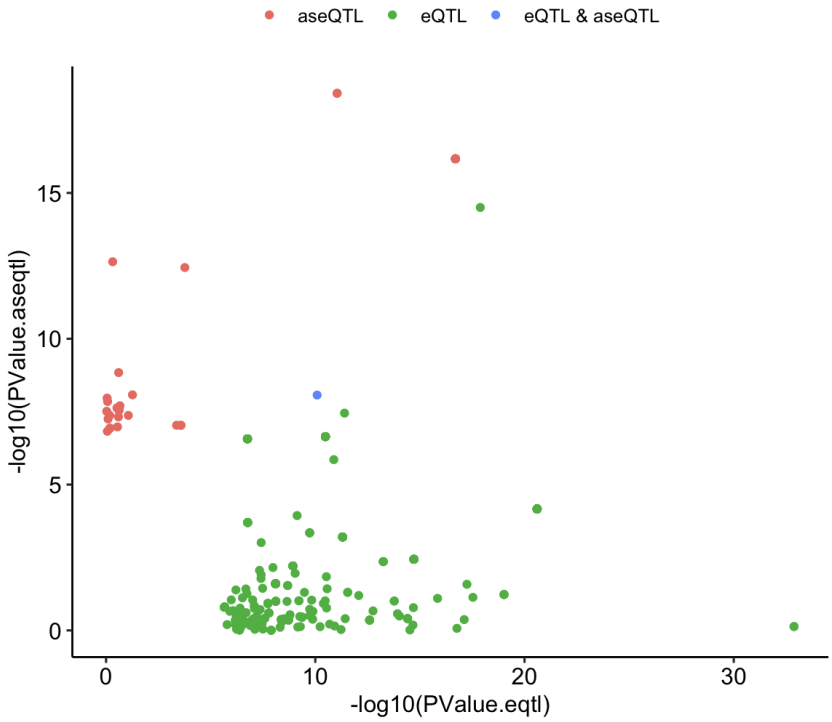
Supplementary figure 7: Significant pathways (FDR < 5%) for gene set enrichment analysis of reactome pathways in genes ranked by mean absolute LogR.



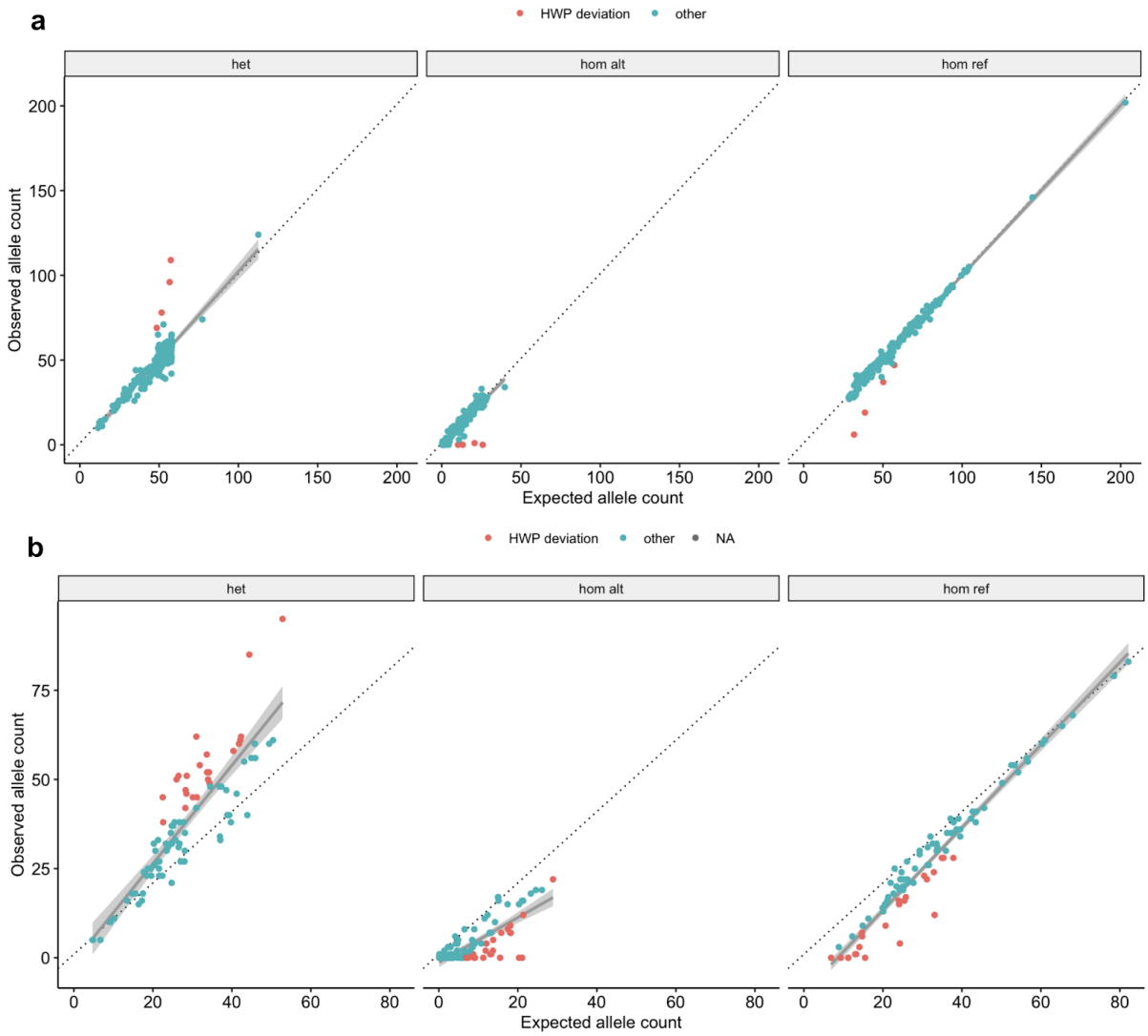
Supplementary figure 8: Reactome pathways enriched in genes differentially expressed for survival status (0.05 FDR, Benjamini-Hochberg). NES: Normalized enrichment score.



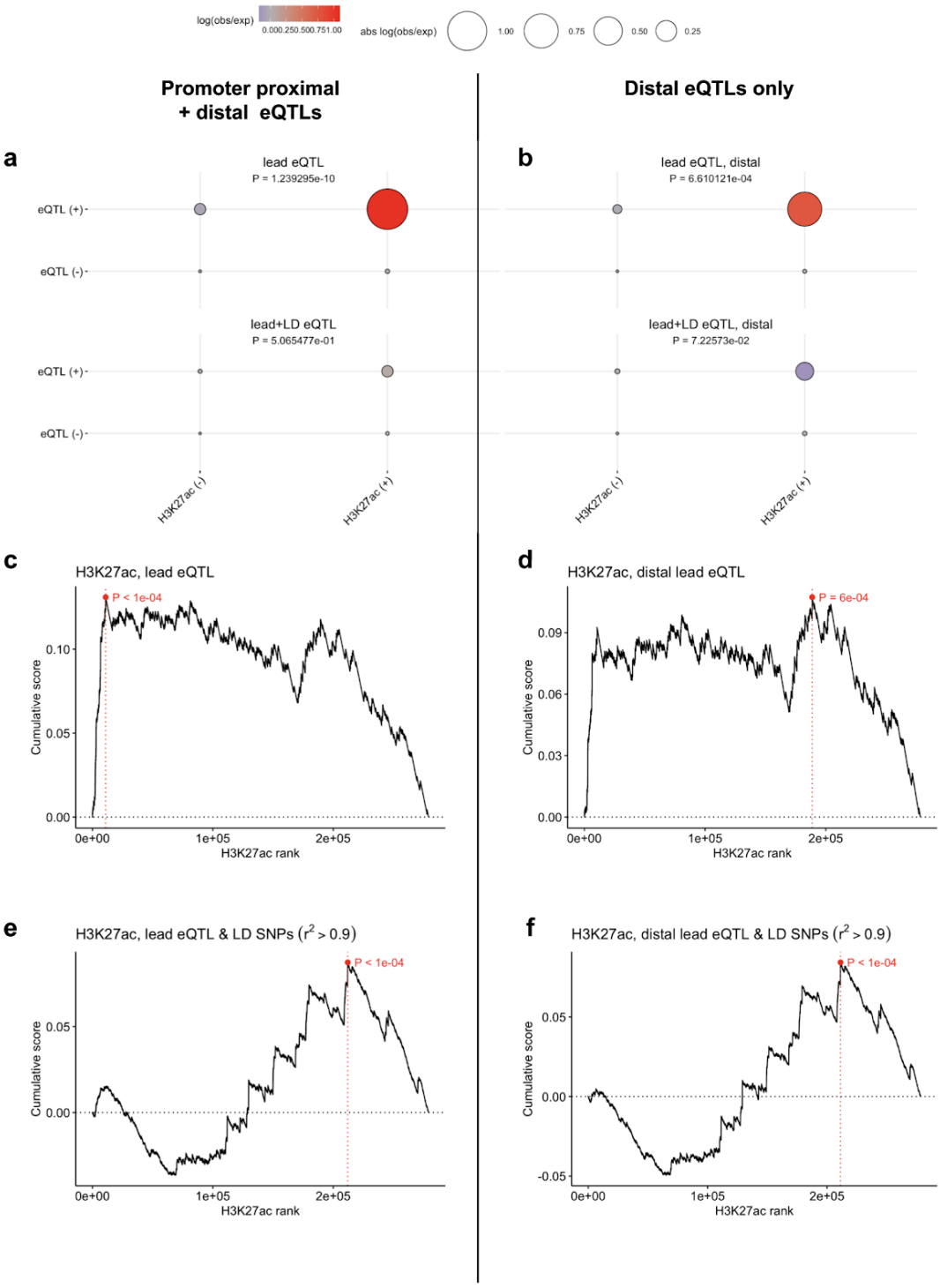
Supplementary figure 9: Allele-specific expression quantitative trait loci associations. **a**, aseQTL associations by distance from gene transcription start site (TSS). Top: Density of TSS distance. Bottom: aseQTL association p-value by TSS distance. Association tests of SNPs > 1,000 kb TSS distance are not shown. If there are multiple lead aseQTL SNPs then they are connected by a red line. Gene's 5' to 3' direction indicated below the plot. **b**, ASE by genotype of lead aseQTL SNP for associated aseQTL genes. For genes with multiple lead aseQTL SNPs and arbitrary one if selected.



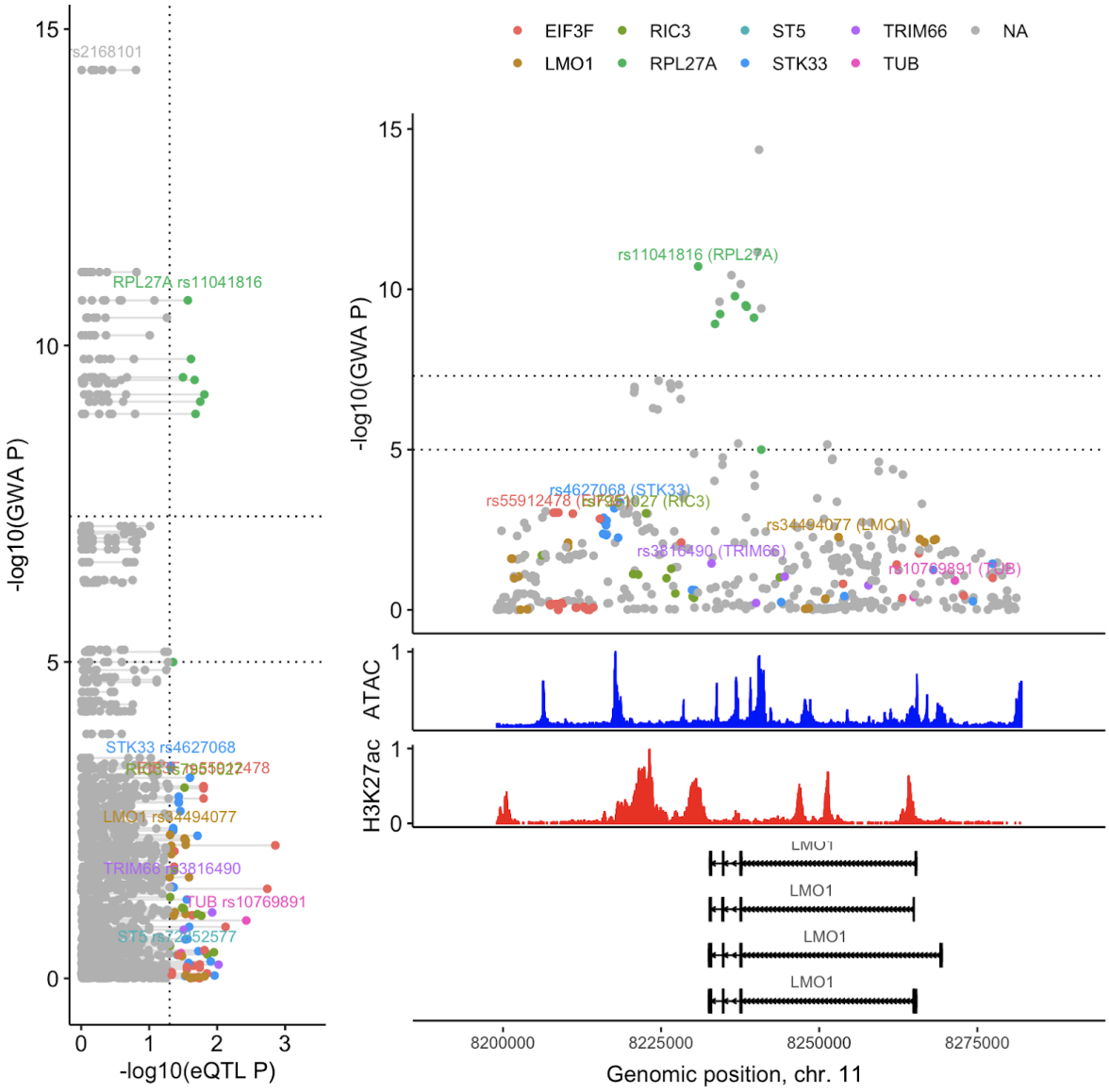
Supplementary figure 10: Comparison between eQTL and aseQTLassociation p-values in lead QTLs. A point corresponds to a test (SNP–gene pair). Only tests of SNPs that are either lead eQTLs or lead aseQTLs are shown (N=452). Colors indicate in which test (eQTL or aseQTL mapping) was found significant.



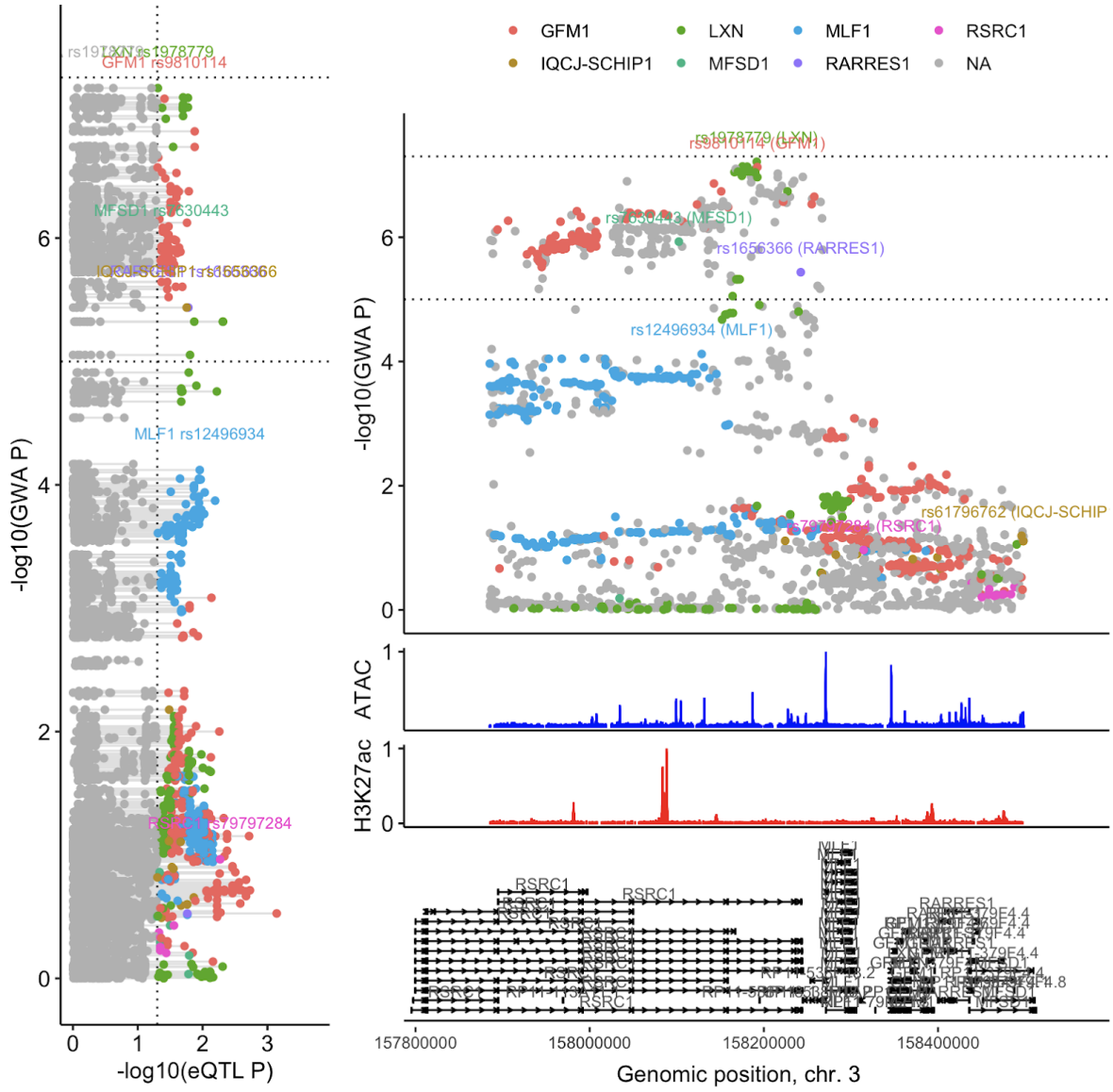
Supplementary figure 11: Genotypes at lead eQTL SNPs. Observed and expected allele counts by genotype for all samples considered in (**a**) eQTL analysis and (**b**) aseQTL analysis. SNPs whose genotypes in informative samples significantly deviate from the Hardy-Weinberg principle (HWP) in red. HWP deviation determined by Chi-square test and significant deviation determined at FDR < 0.05 (Benjamini-Hochberg). hom, homozygous; het, heterozygous; ref: reference allele; alt, alternative allele.



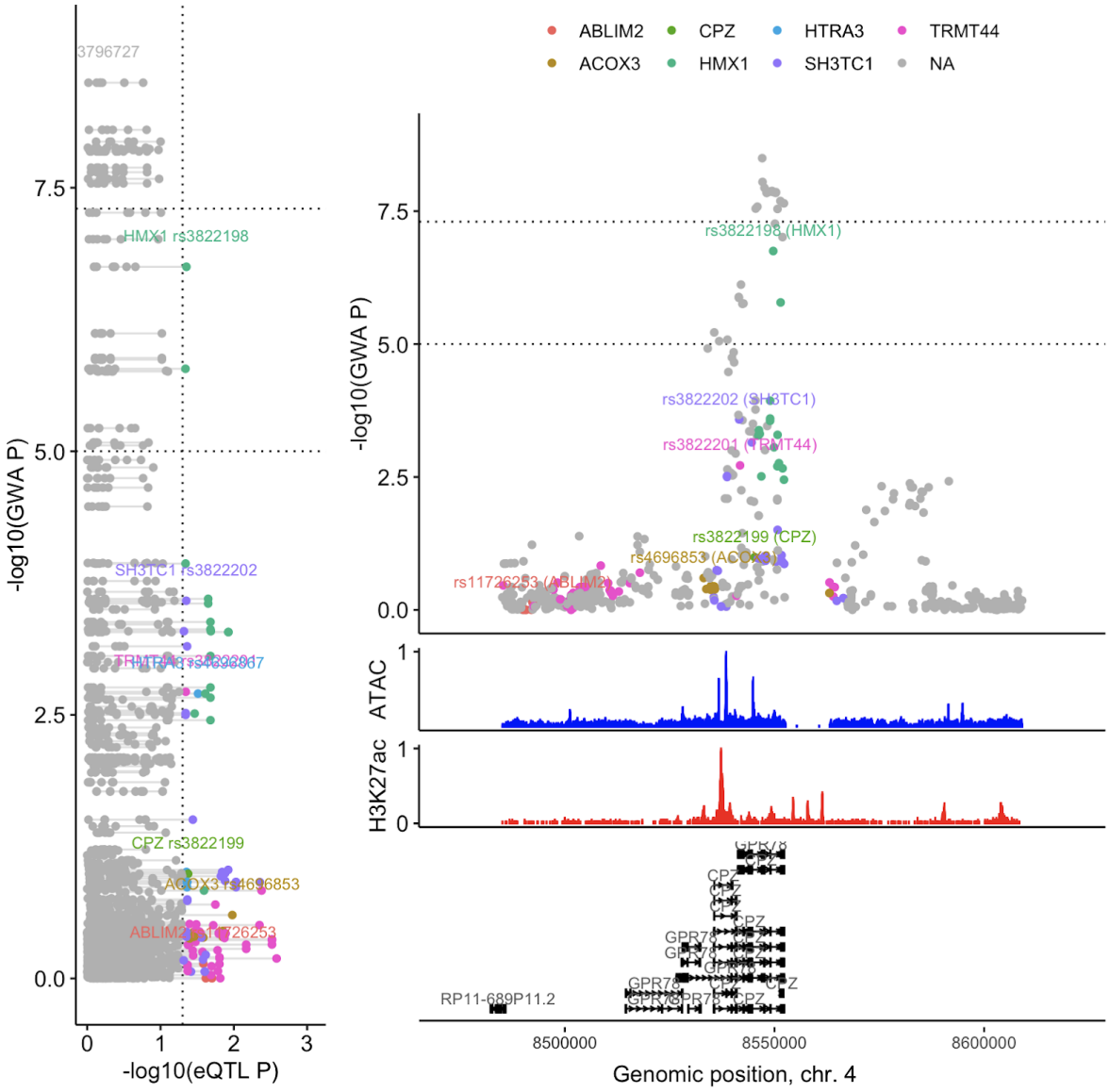
Supplementary figure 13: H3K27ac ChIP-seq features at eQTLs. Lead eQTLs and SNPs in strong LD (r2 > 0.9) with lead eQTLs (LD SNPs) are considered. Overlap of H3K27ac ChIP-seq peaks with lead eQTLs (top) as well as overlap with lead eQTLs and LD SNPs (bottom) for all (**a**) and distal locations only (**b**). ChIP-seq signal enrichment in lead eQTLs for all (**c**) and distal SNPs only (**d**). H3K27ac ChIP-seq signal enrichment in lead eQTLs and LD SNPs for all (**e**) and distal SNPs only **(f)**. p-value in (a,b) obtained by Chi-squared test. p-value in (c-f) obtained by permutation test. Distal SNPs are within +2,000 to -500 bp TSS distance. Maximum cumulative score, its rank and corresponding p-value in (c-f) indicated in red .



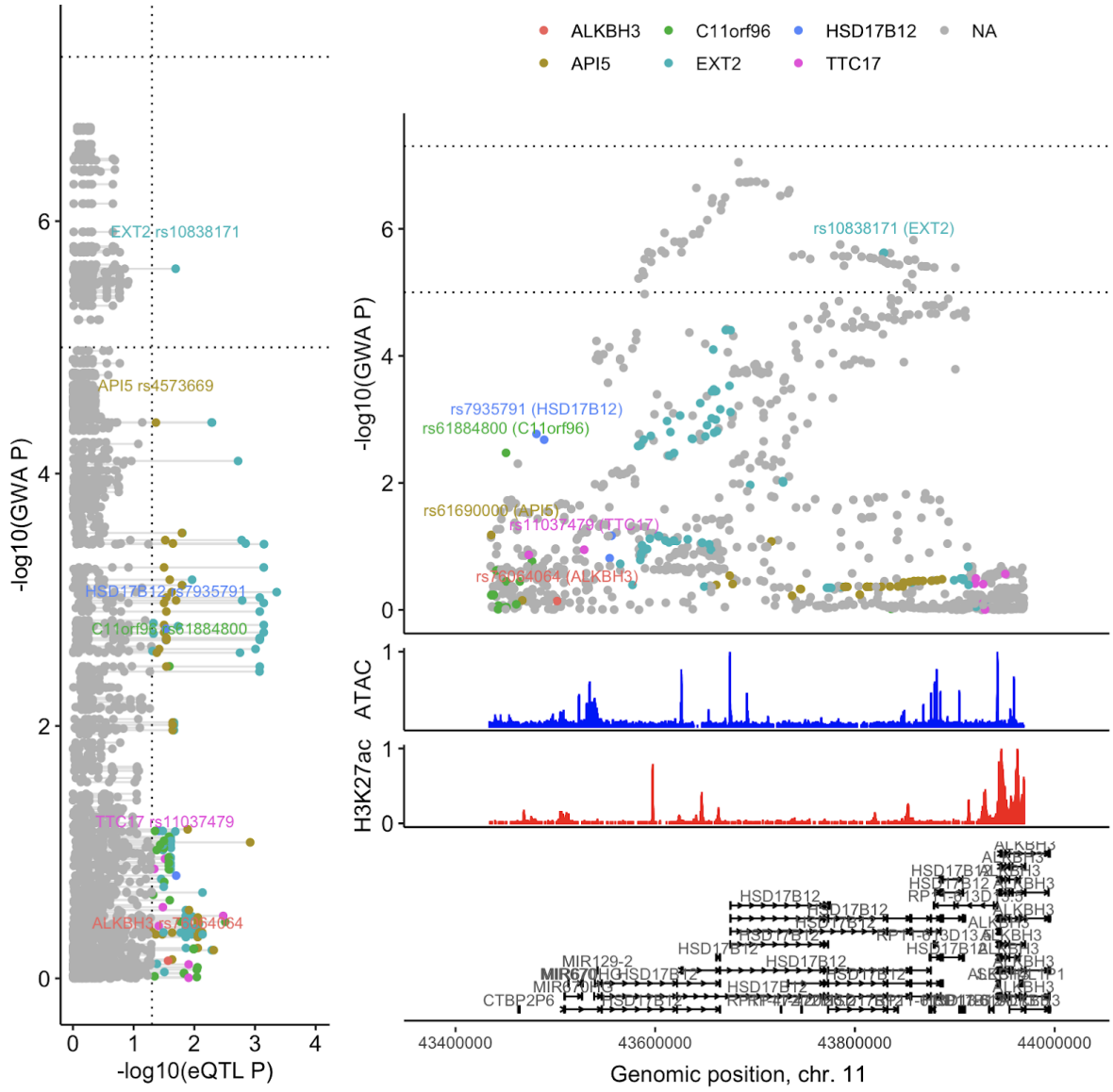
Supplementary figure 12: eQTL analysis at LMO1 risk locus. Left: GWA and eQTL association p-values. eQTL tests of different genes for the same SNP are connected by a grey line. Tests below nominal p-value threshold are color-coded by gene name, others in grey. Dotted line indicates threshold P = 0.05 (nominal). Right: Genome-wide risk association of SNPs from McDaniel et al. 2017. SNPs are annotated by gene with strongest eQTL association with P < 0.05 (nominal). SNPs without eQTL tests and those without tests P < 0.05 (nominal) in grey. H3K27ac-seq ChIP and ATAC-seq signals by read coverage in neuroblastoma cell line SH-SY5Y relative to maximum coverage at the locus.



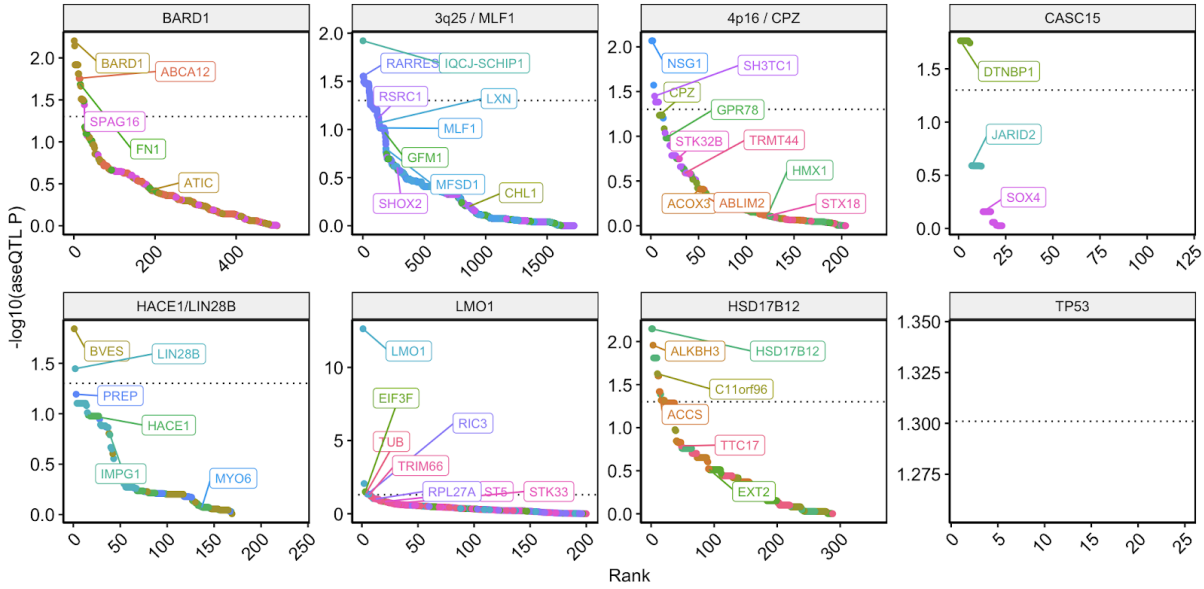
Supplementary figure 14: eQTL analysis at 3q25 / MLF1 risk locus. Left: GWA and eQTL association p-values. eQTL tests of different genes for the same SNP are connected by a grey line. Tests below nominal p-value threshold are color-coded by gene name, others in grey. Dotted line indicates threshold P = 0.05 (nominal). Right: Genome-wide risk association of SNPs from McDaniel et al. 2017. SNPs are annotated by gene with strongest eQTL association with P < 0.05 (nominal). SNPs without eQTL tests and those without tests P < 0.05 (nominal) in grey. H3K27ac-seq ChIP and ATAC-seq signals by read coverage in neuroblastoma cell line SH-SY5Y relative to maximum coverage at the locus.



Supplementary figure 15: eQTL analysis at 4p16/CPZ risk locus. Left: GWA and eQTL association p-values. eQTL tests of different genes for the same SNP are connected by a grey line. Tests below nominal p-value threshold are color-coded by gene name, others in grey. Dotted line indicates threshold P = 0.05 (nominal). Right: Genome-wide risk association of SNPs from McDaniel et al. 2017. SNPs are annotated by gene with strongest eQTL association with P < 0.05 (nominal). SNPs without eQTL tests and those without tests P < 0.05 (nominal) in grey. H3K27ac-seq ChIP and ATAC-seq signals by read coverage in neuroblastoma cell line SH-SY5Y relative to maximum coverage at the locus.



Supplementary figure 16: eQTL analysis at HSD17B12 risk locus. Left: GWA and eQTL association p-values. eQTL tests of different genes for the same SNP are connected by a grey line. Tests below nominal p-value threshold are color-coded by gene name, others in grey. Dotted line indicates threshold P = 0.05 (nominal). Right: Genome-wide risk association of SNPs from McDaniel et al. 2017. SNPs are annotated by gene with strongest eQTL association with P < 0.05 (nominal). SNPs without eQTL tests and those without tests P < 0.05 (nominal) in grey. H3K27ac-seq ChIP and ATAC-seq signals by read coverage in neuroblastoma cell line SH-SY5Y relative to maximum coverage at the locus.



Supplementary figure 17: aseQTL association tests at GWAS risk loci. Color indicates tested gene. Smallest aseQTL association p-value per gene labeled by gene name. Dotted line indicates eQTL association p-value 0.05 (nominal). Only associations of SNPs informative for both GWA and aseQTL analysis shown.

Appendix B: Supplementary tables

Supplementary table 1: Donor sequencing dataset overview.

This table is part of the digital supplementary material.
Filename: suppl_donor_sequencing_dataset_overview.tsv

Supplementary table 2: Manually adjusted purity and ploidy values of samples.

Sample ID	Purity	Ploidy
CB2002	0.80	5.30
CB2006	0.90	2.50
CB2007	0.80	2.95
CB2012	0.55	3.60
CB2014	0.80	3.60
CB2015	0.85	2.50
CB2021	0.88	3.15
CB2024	0.68	3.75
CB2027	0.70	4.00
CB2041	0.95	2.00
CB2042	0.85	5.50
CB2046	0.90	2.90
CB2059	0.85	3.40
NBL15	0.95	3.15
NBL16	0.95	3.30
NBL17	0.95	2.65
NBL22	0.75	3.25
NBL45	0.95	4.50

Supplementary table 3: **Mutect2 command line parameters for somatic SNV calling.**

This table is part of the digital supplementary material.

Filename: suppl_mutect2_command.txt

Supplementary table 4: **FilterMutectCalls command line parameters for somatic SNV calling.**

This table is part of the digital supplementary material.

Filename: suppl_filter_mutect_calls_command.txt

Supplementary table 5: **Candidate gene amplifications identified in 116 neuroblastoma tumor samples.**

Candidate gene amplifications identified in neuroblastoma 116 tumor samples. expr_perc_within_gene: Within-gene expression percentile: 100 corresponds to highest gene expression across samples and 0 to lowest expression.

This table is part of the digital supplementary material.

Filename: suppl_gene_amplifications.tsv

Supplementary table 6: **Gene identifiers of human imprinted genes.**

List of Ensembl gene identifiers of human imprinted genes based on Morison et al. 2005 table 1 and mentions of imprinting in Entrez gene summaries.

This table is part of the digital supplementary material.

Filename: *suppl_imprinted_genes.txt*

Supplementary table 7: **Results of ASE variance component analysis.**

Variance components tumor_purity, cn_ratio_purity, eQTL_het, aseQTL_het, promoter, gene_burden, struct_vars_all, MNA, ase_log_total_count and Residuals. p-values in corresponding columns with pval_* prefix.

This table is part of the digital supplementary material.

Filename: suppl_ase_var_comp_analysis.tsv

Supplementary table 8: **Results of expression variance component analysis.**

Variance components tumor_purity, gene_tumorLogR, eQTL_gt, aseQTL_gt, promoter, gene_burden, struct_vars_all, MNA and Residuals. p-values in corresponding columns with pval_* prefix.

This table is part of the digital supplementary material.

Filename: suppl_expr_var_comp_analysis.tsv

Supplementary table 9: **Reactome pathway enrichment for amplified genes.**

Reactome pathway enrichment for amplified genes (top 30 pathways). p-value derived from a one sided Fisher's exact test. FDR obtained by the Benjamini-Hochberg procedure.

Pathway	p value	FDR
Hydroxycarboxylic acid-binding receptors	1.13×10^{6}	2.52×10^{3}
Keratan sulfate/keratin metabolism	4.64×10^{4}	0.52
Keratan sulfate biosynthesis	1.91×10^{3}	1.00
Regulation of ornithine decarboxylase (ODC)	2.39×10^{3}	1.00
Ubiquitin-dependent degradation of Cyclin D	2.62×10^{3}	1.00
Tyrosine catabolism	3.83×10^{3}	1.00
Cell Cycle	4.07×10^{3}	1.00
Metabolism of polyamines	4.75×10^{3}	1.00
SCF(Skp2)-mediated degradation of p27/p21	5.54×10^{3}	1.00
PTK6 Regulates Cell Cycle	5.66×10^{3}	1.00
Transcriptional regulation by RUNX3	1.03×10^{2}	1.00
Glycosaminoglycan metabolism	1.09×10^{2}	1.00
G1/S Transition	1.45×10^{2}	1.00
Metabolism of RNA	1.87×10^{2}	1.00
Mucopolysaccharidoses	1.94×10^{2}	1.00
Phenylalanine and tyrosine metabolism	1.94×10^{2}	1.00
tRNA processing	1.96×10^{2}	1.00
Defective AVP [...]	2.00×10^{2}	1.00

Defective CYP27B1 [...]	2.00×10^2	1.00
Defective SLC4A1 [...]	2.00×10^2	1.00
MPS IIIB - Sanfilippo syndrome B	2.00×10^2	1.00
MPS IIID - Sanfilippo syndrome D	2.00×10^2	1.00
Severe congenital neutropenia type 4 (G6PC3)	2.00×10^2	1.00
Hh mutants [...] are degraded by ERAD	2.00×10^2	1.00
Regulation of RUNX3 expression and activity	2.00×10^2	1.00
Stabilization of p53	2.13×10^2	1.00
Cyclin E associated events during G1/S transition	2.20×10^2	1.00
p130Cas linkage to MAPK signaling for integrins	2.30×10^2	1.00
Meiosis	2.31×10^2	1.00
Endosomal Sorting Complex Required For Transport (ESCRT)	2.35×10^2	1.00

Supplementary table 10: **Reactome pathways enriched for copy-number dosage effects on gene expression in neuroblastoma tumors.**

This table is part of the digital supplementary material.

Filename: suppl_cn_dosage_effect_sign_pathways.tsv

Supplementary table 11: Association test result between survival status “deceased” and copy-number ratio per chromosome arm.

FWER, family wise error rate as determined by adjusting the ANOVA p-value using the Bonferroni method.

Chromosome arm	Model estimate	ANOVA P	FWER
17p	1.564974132	6.89E-08	2.75E-06
6q	1.254564855	0.01123602966	0.4494411864
6p	1.296267129	0.01203245219	0.4812980878
1p	0.7654859071	0.02165661306	0.8662645224
3q	0.665143573	0.03073428221	1
1q	0.8375154408	0.04037711421	1
8p	0.6947982214	0.04720769673	1
11p	0.4790396507	0.06320542889	1
21p	0.9034793865	0.07360411014	1
2p	0.6729280518	0.1260938228	1

16p	-0.5577954846	0.1766999678	1
5p	0.5406005232	0.1891084218	1
4p	-0.3103390584	0.215511312	1
7q	-0.4855432684	0.2656176117	1
3p	-0.2381615862	0.2986698063	1
4q	0.3421028924	0.3250741053	1
19p	-0.3367052659	0.3274790528	1
2q	0.3601602444	0.3519169018	1
8q	0.2478787628	0.4184387533	1
7p	-0.3296364589	0.4266246079	1
13q	0.258264565	0.5028351473	1
22q	0.2440542707	0.5275962997	1
9q	-0.2170167315	0.5303041429	1
5q	0.3079845765	0.5505798269	1
21q	0.1329354179	0.6334166513	1
18p	-0.1804711371	0.6409305268	1
19q	-0.1633283742	0.6432306724	1
18q	0.1632784252	0.6607003654	1
16q	-0.1562711139	0.7163067131	1
9p	0.1008064359	0.7188943566	1
11q	0.06925833513	0.7591547065	1
14q	0.06660960806	0.8252694049	1
10q	-0.05086346975	0.8461545655	1
15q	0.05372894846	0.8940402367	1
17q	-0.05767210144	0.901306101	1
12p	-0.05418155884	0.9089140487	1
20p	-0.04803526401	0.915112568	1
12q	0.02952760195	0.9486154755	1
10p	0.0147616344	0.9543886766	1
20q	-0.006297667945	0.9904575355	1

Supplementary table 12: Association test result between telomere length ratio (log) and tumor/normal coverage log ratio (logR) per chromosome arm.

FWER, family wise error rate as determined by adjusting the ANOVA p-value using the Bonferroni method.

Chromosome arm	Model estimate	ANOVA P	FWER
11q	-10.48136098	0.0001482765061	0.005931060245
14q	-6.867565622	0.003383794127	0.1353517651
7p	-5.201945842	0.008388792522	0.3355517009
6p	-8.106597936	0.003598575788	0.1439430315
21p	-2.965710544	0.08044587281	1
2q	-4.196442829	0.111609427	1
16p	-7.342112627	0.1262368593	1
5q	-2.811618032	0.1426432062	1
17q	3.259005934	0.1000112608	1
19p	3.065233354	0.1378680321	1
16q	-3.838914245	0.1487226563	1
21q	-2.7248358	0.1436539029	1
10p	-3.657760845	0.2029432693	1
11p	-2.200095667	0.2416785791	1
3p	-1.994229124	0.2462880062	1
17p	1.970755652	0.2688672842	1
1q	-2.442497066	0.3040537263	1
3q	-2.211041166	0.3341654767	1
18q	1.101939704	0.3779759776	1
9p	-1.88753786	0.3897303092	1
13q	-1.435889407	0.4286266148	1
8q	-1.849180668	0.4399740732	1
4p	-1.072628728	0.456008719	1
20p	-1.336838879	0.5560121162	1
1p	-2.914082969	0.5547065839	1
15q	-1.306301829	0.5612464249	1
8p	-1.139152715	0.5831805916	1
6q	1.024580312	0.5915662188	1
4q	-0.8335475151	0.6228142301	1
12p	-0.8693805401	0.6316699707	1
9q	-1.364403795	0.6307622956	1
7q	-0.6569472866	0.6443856394	1
12q	0.6936808624	0.7223965248	1
10q	-0.7512189333	0.7445153944	1
22q	-0.6029557336	0.7606043978	1

18p	0.3090068684	0.7883777136	1
2p	-0.4704288397	0.8067198463	1
19q	0.3822018767	0.8414902748	1
20q	0.2905644218	0.9097596414	1
5p	-0.1737295385	0.9249051832	1

Supplementary table 13: Results of differential gene expression analysis for disease-specific survival.

DEseq2 differential expression (tumor RNA-seq counts) between donors annotated as “deceased from disease” and those with other survival status as defined in the clinical annotation.

This table is part of the digital supplementary material.

Filename: suppl_diff_expr_deceased.tsv

Supplementary table 14: Function network enrichment (GO biological processes) of differentially expressed genes (disease-specific survival) on 17p that show significant CN dosage effects (STRING DB).

Term ID	Term description	FDR
GO:0051129	negative regulation of cellular component organization	0.0028
GO:0001941	postsynaptic membrane organization	0.0329
GO:0007268	chemical synaptic transmission	0.0329
GO:0007626	locomotory behavior	0.0329
GO:0008104	protein localization	0.0329
GO:0010975	regulation of neuron projection development	0.0329
GO:0010977	negative regulation of neuron projection development	0.0329
GO:0033036	macromolecule localization	0.0329
GO:0046328	regulation of JNK cascade	0.0329
GO:0050885	neuromuscular process controlling balance	0.0329
GO:0051128	regulation of cellular component organization	0.0329
GO:0061024	membrane organization	0.0329
GO:0099601	regulation of neurotransmitter receptor activity	0.0329
GO:0150012	positive regulation of neuron projection arborization	0.0329
GO:1900449	regulation of glutamate receptor signaling pathway	0.0329

GO:2001257	regulation of cation channel activity	0.0329
GO:0035641	locomotory exploration behavior	0.0466
GO:0007610	behavior	0.0496

Supplementary table 15: Results of differential expression test between samples with and without alternative lengthening of telomeres (ALT) phenotype and gene expression correlation with 11q logR.

Pearson correlation coefficient between expression residual of tested genes and 11q logR across samples in column “cor_logr_11q”.

This table is part of the digital supplementary material.

Filename: suppl_diff_expr_ALT_cor_logr_11q.tsv

Supplementary table 16: Allelic regulated (AR) genes, in which ASE and total gene expression is correlated.

Effect sizes and statistics of AR test in columns effect_size, fit_P, fit_P.adj. Differential expression test results in columns: log2FoldChange, pval, padj

This table is part of the digital supplementary material.

Filename: suppl_AR_genes.tsv

Supplementary table 17: eQTL and aseQTL association test results.

eQTL and aseQTL association p-values and effect sizes in columns prefixed by PValue.* and Z.* respectively.

This table is part of the digital supplementary material.

Filename: suppl_qtl_results.tsv.gz

Supplementary table 18: Candidate regulatory SNPs for identified eQTL genes prioritized bei LD with lead eQTL SNP and overlap with ATAC peaks identified in neuroblastoma cell line SH-SY5Y.

Gene name	RSID	eQTL P	Lead eQTL P	TSS distance	H3K27ac peak
ACCS	rs2074038	8.17E-11	8.17E-11	514	TRUE
ADHFE1	rs2433593	5.99E-06	2.15E-06	9347	FALSE
AGA	rs11131799	3.87E-07	3.87E-07	279	TRUE
ANAPC4	rs3822217	6.88E-06	5.11E-07	-116	TRUE
APOL2	rs5756111	4.64E-08	4.64E-08	244	FALSE
ARL17A	rs2316951	7.93E-09	1.82E-09	518005	FALSE
ARL17A	rs529556708	7.93E-09	1.82E-09	385658	FALSE
ARL17A	rs553226241	7.93E-09	1.82E-09	384536	FALSE
ARL17A	rs575200428	7.93E-09	1.82E-09	535171	FALSE
ARL17A	rs111413387	8.67E-09	1.82E-09	312679	FALSE
ARL17A	rs554959222	1.20E-08	1.82E-09	312632	TRUE
ARL17A	rs113417378	1.22E-08	1.82E-09	386279	FALSE
ARPC5L	rs10760379	1.71E-07	3.51E-08	-9177	FALSE
ARPC5L	rs10986466	3.59E-07	3.51E-08	-8865	FALSE
ARPC5L	rs10760380	5.31E-07	3.51E-08	-8086	FALSE
ARPC5L	rs12375547	5.31E-07	3.51E-08	-7872	FALSE
C17orf97	rs11150882	7.84E-17	7.84E-17	-470	TRUE
C17orf97	rs7502594	7.84E-17	7.84E-17	64	FALSE
C17orf97	rs7503725	7.84E-17	7.84E-17	24	FALSE
CBLN3	rs4344657	5.88E-07	1.83E-08	796	TRUE
CCDC163P	rs3748643	1.77E-10	2.71E-11	48	FALSE
CD46	rs6540443	8.73E-06	3.94E-07	76608	FALSE
CDC7	rs13447455	2.58E-08	2.58E-08	37	TRUE
CHURC1	rs72726294	1.15E-09	7.40E-10	-57	FALSE
CLDN4	rs6946037	2.41E-06	1.64E-06	22843	FALSE
CLDN4	rs6949053	2.67E-06	1.64E-06	56065	FALSE
CLDN4	rs10276377	4.73E-06	1.64E-06	42966	FALSE
CLDN4	rs10233067	4.02E-05	1.64E-06	43048	FALSE
CYP4V2	rs7663027	4.11E-11	4.11E-11	-45	TRUE
CYP4V2	rs2241819	1.91E-09	4.11E-11	152	TRUE
DCXR	rs57552134	5.58E-07	5.58E-07	-8	FALSE
DNAJC15	rs12015	3.29E-11	3.29E-11	526	TRUE
DNAJC15	rs2281777	3.29E-11	3.29E-11	593	TRUE
DNAJC15	rs2281778	3.29E-11	3.29E-11	638	TRUE
DNAJC15	rs17553284	4.97E-11	3.29E-11	367	TRUE

DNAJC15	rs2281780	4.97E-11	3.29E-11	656	TRUE
DNAJC15	rs2281779	1.04E-10	3.29E-11	641	TRUE
EFCAB2	rs61844237	2.96E-11	2.96E-11	655	TRUE
EFHB	rs11128927	1.33E-10	1.33E-10	80	TRUE
EXOSC6	rs4985407	1.69E-07	4.91E-08	-68	FALSE
FAHD1	rs28364709	2.33E-09	6.88E-10	207	TRUE
FAHD1	rs2369275	7.54E-09	6.88E-10	45344	FALSE
FAHD1	rs62038433	7.54E-09	6.88E-10	19052	FALSE
FAHD1	rs62038434	7.54E-09	6.88E-10	19056	FALSE
FAHD1	rs62038435	7.54E-09	6.88E-10	19133	FALSE
FAHD1	rs62038436	7.54E-09	6.88E-10	19369	FALSE
GBP3	rs12127787	3.52E-17	5.66E-18	29816	FALSE
GLIPR1L2	rs11180483	5.17E-14	3.00E-15	250	TRUE
GLIPR1L2	rs7978856	5.17E-14	3.00E-15	-77	FALSE
GNPDA2	rs5007781	2.85E-08	1.65E-08	-33765	FALSE
GNRHR	rs28653581	1.18E-10	2.89E-11	53053	FALSE
GNRHR	rs78578320	1.18E-10	2.89E-11	53389	TRUE
KANSL1	rs111413387	8.69E-09	1.49E-10	-41676	FALSE
KANSL1	rs554959222	9.02E-09	1.49E-10	-41723	TRUE
KANSL1	rs2316951	1.42E-08	1.49E-10	163650	FALSE
KANSL1	rs529556708	1.42E-08	1.49E-10	31303	FALSE
KANSL1	rs553226241	1.42E-08	1.49E-10	30181	FALSE
KANSL1	rs575200428	1.42E-08	1.49E-10	180816	FALSE
KANSL1	rs62063779	1.42E-08	1.49E-10	248062	FALSE
KANSL1	rs62064663	1.42E-08	1.49E-10	222694	FALSE
KANSL1	rs62064664	1.42E-08	1.49E-10	221271	FALSE
KANSL1	rs62064665	1.42E-08	1.49E-10	221206	FALSE
KANSL1	rs76618565	1.42E-08	1.49E-10	234252	FALSE
KANSL1	rs77290642	1.42E-08	1.49E-10	234253	FALSE
KANSL1	rs113417378	2.00E-08	1.49E-10	31924	FALSE
KAT8	rs8050894	7.72E-06	5.66E-07	-22566	FALSE
KIAA1143	rs2279908	7.94E-09	1.42E-09	32026	TRUE
LAMB3	rs56071308	4.57E-06	1.46E-06	-183593	FALSE
LAMB3	rs56268268	4.57E-06	1.46E-06	-183573	FALSE
LCA5L	rs13051142	2.35E-11	5.89E-12	301	TRUE
LCA5L	rs13050837	2.68E-11	5.89E-12	375	TRUE
LCA5L	rs9983716	2.68E-11	5.89E-12	791	FALSE
LCA5L	rs13051054	3.44E-11	5.89E-12	224	TRUE
LRRC23	ss1388026427	7.38E-07	2.13E-07	31343	TRUE
LRRC23	ss1388026429	7.38E-07	2.13E-07	31374	TRUE

LRRC23	ss1388026727	1.93E-06	2.13E-07	41071	TRUE
LRRC37A2	rs2316951	1.21E-10	1.60E-11	-449794	FALSE
LRRC37A2	rs529556708	1.21E-10	1.60E-11	-317447	FALSE
LRRC37A2	rs553226241	1.21E-10	1.60E-11	-316325	FALSE
LRRC37A2	rs575200428	1.21E-10	1.60E-11	-466960	FALSE
LRRC37A2	rs111413387	2.85E-10	1.60E-11	-244468	FALSE
LRRC37A2	rs554959222	3.46E-10	1.60E-11	-244421	TRUE
LRRC37A2	rs113417378	4.14E-10	1.60E-11	-318068	FALSE
LRRIQ3	rs11806946	2.28E-07	1.64E-07	162	TRUE
LRRIQ3	rs1412825	2.28E-07	1.64E-07	57	FALSE
LSG1	rs7619357	1.18E-09	1.18E-09	-13301	FALSE
LYZ	rs2249093	1.75E-07	1.75E-07	11391	FALSE
LYZ	rs2617835	1.75E-07	1.75E-07	11299	FALSE
LYZ	rs634512	1.75E-07	1.75E-07	11891	FALSE
LYZ	rs554591	2.56E-07	1.75E-07	11726	TRUE
LYZ	rs623853	2.56E-07	1.75E-07	11709	TRUE
MASTL	rs7919803	2.80E-09	1.99E-09	575	TRUE
MASTL	rs3824593	8.10E-09	1.99E-09	550	TRUE
MASTL	rs10829181	1.89E-08	1.99E-09	-54205	TRUE
MERTK	rs72825673	7.75E-09	7.75E-09	128026	TRUE
METTL21B	rs10747783	4.23E-10	4.23E-10	11339	FALSE
MTRF1L	rs3818127	1.01E-12	2.54E-13	-76	TRUE
MTRF1L	rs3818130	1.01E-12	2.54E-13	-336	TRUE
NDUFS5	rs3768324	1.43E-07	5.45E-08	472	TRUE
NSA2	rs79669494	7.92E-09	7.92E-09	451	TRUE
OCLN	rs8192259	1.74E-06	1.74E-06	-325318	FALSE
OMA1	rs1109895	4.30E-07	6.29E-08	43	FALSE
OMA1	rs1109896	4.30E-07	6.29E-08	70	FALSE
OMA1	rs2087799	4.30E-07	6.29E-08	550	TRUE
OMA1	rs2406784	4.30E-07	6.29E-08	487	TRUE
OMA1	rs2406785	4.30E-07	6.29E-08	489	TRUE
PPIL3	rs7559150	4.12E-18	2.93E-18	-37	FALSE
RBL2	rs6499613	3.71E-11	2.83E-12	-4471	FALSE
RBL2	rs8055642	6.77E-11	2.83E-12	69692	FALSE
ROPN1B	rs12636284	6.36E-10	1.81E-11	21226	TRUE
ROPN1B	rs35120077	6.36E-10	1.81E-11	21379	TRUE
RP11-166B2.1	rs393329	9.94E-27	9.94E-27	78	TRUE
RPS26	rs10876864	1.31E-33	1.31E-33	-34552	FALSE
RPS26	rs1131017	1.31E-33	1.31E-33	292	TRUE
RPS26	rs7297175	3.57E-33	1.31E-33	38171	FALSE

SETD9	rs185220	4.88E-08	3.86E-08	270	TRUE
SNX16	rs4316108	8.81E-09	3.07E-09	736	FALSE
SNX16	rs10097100	1.01E-08	3.07E-09	249	FALSE
SNX19	rs7936858	3.17E-07	3.95E-08	54304	FALSE
STAT4	rs13019004	2.94E-10	1.30E-10	-93897	TRUE
STAT4	rs34765012	2.94E-10	1.30E-10	-96330	FALSE
STAT4	rs4146105	2.94E-10	1.30E-10	-93644	FALSE
TAF1B	rs2303914	6.06E-10	6.06E-10	203	TRUE
U2AF1L4	rs2293686	2.07E-07	8.13E-08	418	FALSE
U2AF1L4	rs3746277	2.07E-07	8.13E-08	-3138	FALSE
WBSCR27	rs6949053	3.22E-10	3.22E-10	-13072	FALSE
WBSCR27	rs6946037	3.77E-10	3.22E-10	20150	FALSE
WBSCR27	rs10276377	9.48E-10	3.22E-10	27	FALSE
WBSCR27	rs10233067	1.89E-08	3.22E-10	-55	FALSE
XRRA1	rs2165163	3.80E-19	9.54E-20	102	TRUE
XRRA1	rs12421899	5.45E-18	9.54E-20	202199	FALSE
ZNF124	rs10924924	1.90E-09	1.90E-09	51998	FALSE
ZNF266	rs10418910	3.91E-14	9.55E-15	125418	FALSE
ZNF266	rs10411624	5.22E-14	9.55E-15	111324	TRUE
ZNF429	rs117047235	1.79E-06	5.22E-08	89961	TRUE
ZNF429	rs62110212	1.79E-06	5.22E-08	89851	TRUE

Supplementary table 19: Cox proportional hazards regression model results for lead eQTL genotypes and overall survival.

This table is part of the digital supplementary material.
Filename: suppl_lead_eqtl_cox.tsv

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Abe, Masanobu, Miki Ohira, Atsushi Kaneda, Yukiko Yagi, Seiichiro Yamamoto, Yoshihiro Kitano, Tsuyoshi Takato, Akira Nakagawara, and Toshikazu Ushijima. 2005. “CpG Island Methylator Phenotype Is a Strong Determinant of Poor Prognosis in Neuroblastomas.” *Cancer Research* 65 (3): 828–34. https://www.ncbi.nlm.nih.gov/pubmed/15705880.

Abe, Masanobu, Frank Westermann, Akira Nakagawara, Tsuyoshi Takato, Manfred Schwab, and Toshikazu Ushijima. 2007. “Marked and Independent Prognostic Significance of the CpG Island Methylator Phenotype in Neuroblastomas.” *Cancer Letters* 247 (2): 253–58. https://doi.org/10.1016/j.canlet.2006.05.001.

Abraham, Brian J., Denes Hnisz, Abraham S. Weintraub, Nicholas Kwiatkowski, Charles H. Li, Zhaodong Li, Nina Weichert-Leahey, et al. 2017. “Small Genomic Insertions Form Enhancers That Misregulate Oncogenes.” *Nature Communications* 8 (February): 14385. https://doi.org/10.1038/ncomms14385.

Ackermann, Sandra, Maria Cartolano, Barbara Hero, Anne Welte, Yvonne Kahlert, Andrea Roderwieser, Christoph Bartenhagen, et al. 2018. “A Mechanistic Classification of Clinical Phenotypes in Neuroblastoma.” *Science* 362 (6419): 1165–70. https://doi.org/10.1126/science.aat6768.

Ackermann, Sandra, Hayriye Kocak, Barbara Hero, Volker Ehemann, Yvonne Kahlert, André Oberthuer, Frederik Roels, et al. 2014. “FOXP1 Inhibits Cell Growth and Attenuates Tumorigenicity of Neuroblastoma.” *BMC Cancer* 14 (November): 840. https://doi.org/10.1186/1471-2407-14-840.

Adey, Andrew, Hilary G. Morrison, Asan, Xu Xun, Jacob O. Kitzman, Emily H. Turner, Bethany Stackhouse, et al. 2010. “Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density in Vitro Transposition.” *Genome Biology* 11 (12): R119. https://doi.org/10.1186/gb-2010-11-12-r119.

Ahmad, Kami, and Steven Henikoff. 2002. “The Histone Variant H3.3 Marks Active Chromatin by Replication-Independent Nucleosome Assembly.” *Molecular Cell* 9 (6): 1191–1200. https://doi.org/10.1016/s1097-2765(02)00542-7.

Al-Kuraya, Khawla, Peter Schraml, Joachim Torhorst, Coya Tapia, Boriana Zaharieva, Hedvika Novotny, Hanspeter Spichtin, et al. 2004. “Prognostic Relevance of Gene Amplifications and Coamplifications in Breast Cancer.” *Cancer Research* 64 (23): 8534–40. https://doi.org/10.1158/0008-5472.CAN-04-1945.

Allfrey, V. G., R. Faulkner, and A. E. Mirsky. 1964. “ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS.” *Proceedings of the National Academy of Sciences of the United States of America* 51 (May): 786–94. https://doi.org/10.1073/pnas.51.5.786.

Alvarez, Mariano J., Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, and Andrea Califano. 2016. “Functional Characterization of Somatic Mutations in Cancer Using Network-Based Inference of Protein Activity.” *Nature Genetics* 48 (8): 838–47. https://doi.org/10.1038/ng.3593.

Amarasinghe, Kaushalya C., Jason Li, and Saman K. Halgamuge. 2013. “CoNVEX: Copy Number Variation Estimation in Exome Sequencing Data Using HMM.” *BMC Bioinformatics* 14 Suppl 2 (January): S2. https://doi.org/10.1186/1471-2105-14-S2-S2.

Amarasinghe, Kaushalya C., Jason Li, Sally M. Hunter, Georgina L. Ryland, Prue A. Cowin, Ian G. Campbell, and Saman K. Halgamuge. 2014. “Inferring Copy Number and Genotype in Tumour Exome Data.” *BMC Genomics* 15 (August): 732.

https://doi.org/10.1186/1471-2164-15-732.
Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69. https://doi.org/10.1093/bioinformatics/btu638.
Aplan, P. D., D. P. Lombardi, A. M. Ginsberg, J. Cossman, V. L. Bertness, and I. R. Kirsch. 1990. “Disruption of the Human SCL Locus by ‘Illegitimate’ V-(D)-J Recombinase Activity.” *Science* 250 (4986): 1426–29. https://doi.org/10.1126/science.2255914.
Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq.” *Science* 339 (6123): 1074–77. https://doi.org/10.1126/science.1232542.
Arthur, Sarah E., Aixiang Jiang, Bruno M. Grande, Miguel Alcaide, Razvan Cojocaru, Christopher K. Rushton, Anja Mottok, et al. 2018. “Genome-Wide Discovery of Somatic Regulatory Variants in Diffuse Large B-Cell Lymphoma.” *Nature Communications* 9 (1): 4001. https://doi.org/10.1038/s41467-018-06354-3.
Astuti, D., A. Agathanggelou, S. Honorio, A. Dallol, T. Martinsson, P. Kogner, C. Cummins, et al. 2001. “RASSF1A Promoter Region CpG Island Hypermethylation in Phaeochromocytomas and Neuroblastoma Tumours.” *Oncogene* 20 (51): 7573–77. https://doi.org/10.1038/sj.onc.1204968.
Attig, Jan, George R. Young, Louise Hosie, David Perkins, Vesela Encheva-Yokoya, Jonathan P. Stoye, Ambrosius P. Snijders, Nicola Ternette, and George Kassiotis. 2019. “LTR Retroelement Expansion of the Human Cancer Transcriptome and Immunopeptidome Revealed by de Novo Transcript Assembly.” *Genome Research* 29 (10): 1578–90. https://doi.org/10.1101/gr.248922.119.
Awasthi, Anshul, Adele G. Woolley, Fabienne J. Lecomte, Noelyn Hung, Bruce C. Baguley, Sigurd M. Wilbanks, Aaron R. Jeffs, and Joel D. A. Tyndall. 2013. “Variable Expression of GLIPR1 Correlates with Invasive Potential in Melanoma Cells.” *Frontiers in Oncology* 3 (August): 225. https://doi.org/10.3389/fonc.2013.00225.
Baca, Sylvan C., Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, et al. 2013. “Punctuated Evolution of Prostate Cancer Genomes.” *Cell* 153 (3): 666–77. https://doi.org/10.1016/j.cell.2013.03.021.
Bagatell, Rochelle, Maja Beck-Popovic, Wendy B. London, Yang Zhang, Andrew D. J. Pearson, Katherine K. Matthay, Tom Monclair, Peter F. Ambros, Susan L. Cohn, and International Neuroblastoma Risk Group. 2009. “Significance of MYCN Amplification in International Neuroblastoma Staging System Stage 1 and 2 Neuroblastoma: A Report from the International Neuroblastoma Risk Group Database.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (3): 365–70. https://doi.org/10.1200/JCO.2008.17.9184.
Balaban-Malenbaum, G., and F. Gilbert. 1977. “Double Minute Chromosomes and the Homogeneously Staining Regions in Chromosomes of a Human Neuroblastoma Cell Line.” *Science* 198 (4318): 739–41. https://doi.org/10.1126/science.71759.
Bamford, S., E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, et al. 2004. “The COSMIC (Catalogue of Somatic Mutations in Cancer) Database and Website.” *British Journal of Cancer* 91 (2): 355–58. https://doi.org/10.1038/sj.bjc.6601894.
Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell* 129 (4): 823–37. https://doi.org/10.1016/j.cell.2007.05.009.
Barthel, Floris P., Wei Wei, Ming Tang, Emmanuel Martinez-Ledesma, Xin Hu, Samirkumar B. Amin, Kadir C. Akdemir, et al. 2017. “Systematic Analysis of Telomere Length and

Somatic Alterations in 31 Cancer Types." *Nature Genetics* 49 (3): 349–57. https://doi.org/10.1038/ng.3781.
Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, et al. 2014. "Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals." *Genome Research* 24 (1): 14–24. https://doi.org/10.1101/gr.155192.113.
Baudis, Michael. 2007. "Genomic Imbalances in 5918 Malignant Epithelial Tumors: An Explorative Meta-Analysis of Chromosomal CGH Data." *BMC Cancer* 7 (December): 226. https://doi.org/10.1186/1471-2407-7-226.
Bauer, Daniel E., Sophia C. Kamran, Samuel Lessard, Jian Xu, Yuko Fujiwara, Carrie Lin, Zhen Shao, et al. 2013. "An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level." *Science* 342 (6155): 253–57. https://doi.org/10.1126/science.1242088.
Behjati, Sam, Patrick S. Tarpey, Nadège Presneau, Susanne Scheipl, Nischalan Pillay, Peter Van Loo, David C. Wedge, et al. 2013. "Distinct H3F3A and H3F3B Driver Mutations Define Chondroblastoma and Giant Cell Tumor of Bone." *Nature Genetics* 45 (12): 1479–82. https://doi.org/10.1038/ng.2814.
Bell, A. C., and G. Felsenfeld. 2000. "Methylation of a CTCF-Dependent Boundary Controls Imprinted Expression of the Igf2 Gene." *Nature* 405 (6785): 482–85. https://doi.org/10.1038/35013100.
Ben-David, Uri, and Angelika Amon. 2019. "Context Is Everything: Aneuploidy in Cancer." *Nature Reviews. Genetics* 21 (1): 44–62. https://doi.org/10.1038/s41576-019-0171-x.
Bergamaschi, Anna, Young H. Kim, Pei Wang, Therese Sørlie, Tina Hernandez-Boussard, Per E. Lonning, Robert Tibshirani, Anne-Lise Børresen-Dale, and Jonathan R. Pollack. 2006. "Distinct Patterns of DNA Copy Number Alteration Are Associated with Different Clinicopathological Features and Gene-Expression Subtypes of Breast Cancer." *Genes, Chromosomes & Cancer* 45 (11): 1033–40. https://doi.org/10.1002/gcc.20366.
Berger, Alice H., Alfred G. Knudson, and Pier Paolo Pandolfi. 2011. "A Continuum Model for Tumour Suppression." *Nature* 476 (7359): 163–69. https://doi.org/10.1038/nature10275.
Beroukhim, Rameen, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, et al. 2010. "The Landscape of Somatic Copy-Number Alteration across Human Cancers." *Nature* 463 (7283): 899–905. https://doi.org/10.1038/nature08822.
Bielle, Franck, Paul Fréneaux, Corinne Jeanne-Pasquier, Aurélie Maran-Gonzalez, Audrey Rousseau, Laurence Lamant, Régine Paris, et al. 2012. "PHOX2B Immunolabeling: A Novel Tool for the Diagnosis of Undifferentiated Neuroblastomas among Childhood Small Round Blue-Cell Tumors." *The American Journal of Surgical Pathology* 36 (8): 1141–49. https://doi.org/10.1097/PAS.0b013e31825a6895.
Bignell, Graham R., Jing Huang, Joel Greshock, Stephen Watt, Adam Butler, Sofie West, Mira Grigorova, et al. 2004. "High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays." *Genome Research* 14 (2): 287–95. https://doi.org/10.1101/gr.2012304.
Bird, A. P. 1980. "DNA Methylation and the Frequency of CpG in Animal DNA." *Nucleic Acids Research* 8 (7): 1499–1504. https://doi.org/10.1093/nar/8.7.1499.
Bloom, L., and H. R. Horvitz. 1997. "The Caenorhabditis Elegans Gene Unc-76 and Its Human Homologs Define a New Gene Family Involved in Axonal Outgrowth and Fasciculation." *Proceedings of the National Academy of Sciences of the United States of America* 94 (7): 3414–19. https://doi.org/10.1073/pnas.94.7.3414.
Blow, Matthew J., David J. McCulley, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2010. "ChIP-Seq Identification of Weakly Conserved Heart Enhancers." *Nature Genetics* 42 (9): 806–10. https://doi.org/10.1038/ng.650.

Boeva, Valentina, Caroline Louis-Brennetot, Agathe Peltier, Simon Durand, Cécile Pierre-Eugène, Virginie Raynal, Heather C. Etchevers, et al. 2017. “Heterogeneity of Neuroblastoma Cell Identity Defined by Transcriptional Circuitries.” *Nature Genetics* 49 (9): 1408–13. https://doi.org/10.1038/ng.3921.

Boeva, Valentina, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. 2012. “Control-FREEC: A Tool for Assessing Copy Number and Allelic Content Using next-Generation Sequencing Data.” *Bioinformatics* 28 (3): 423–25. https://doi.org/10.1093/bioinformatics/btr670.

Bordow, S. B., M. D. Norris, P. S. Haber, G. M. Marshall, and M. Haber. 1998. “Prognostic Significance of MYCN Oncogene Expression in Childhood Neuroblastoma.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 16 (10): 3286–94. https://doi.org/10.1200/JCO.1998.16.10.3286.

Bosse, Kristopher R., Sharon J. Diskin, Kristina A. Cole, Andrew C. Wood, Robert W. Schnepp, Geoffrey Norris, Le B. Nguyen, et al. 2012. “Common Variation at BARD1 Results in the Expression of an Oncogenic Isoform That Influences Neuroblastoma Susceptibility and Oncogenicity.” *Cancer Research* 72 (8): 2068–78. https://doi.org/10.1158/0008-5472.CAN-11-3703.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. “High-Resolution Mapping and Characterization of Open Chromatin across the Genome.” *Cell* 132 (2): 311–22. https://doi.org/10.1016/j.cell.2007.12.014.

Bragg, D. Cristopher, Kotchaphorn Mangkalaphiban, Christine A. Vaine, Nichita J. Kulkarni, David Shin, Rachita Yadav, Jyotsna Dhakal, et al. 2017. “Disease Onset in X-Linked Dystonia-Parkinsonism Correlates with Expansion of a Hexameric Repeat within an SVA Retrotransposon in TAF1.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (51): E11020–28. https://doi.org/10.1073/pnas.1712526114.

Breit, T. M., E. J. Mol, I. L. Wolvers-Tettero, W. D. Ludwig, E. R. van Wering, and J. J. van Dongen. 1993. “Site-Specific Deletions Involving the Tal-1 and Sil Genes Are Restricted to Cells of the T Cell Receptor Alpha/beta Lineage: T Cell Receptor Delta Gene Deletion Mechanism Affects Multiple Genes.” *The Journal of Experimental Medicine* 177 (4): 965–77. https://doi.org/10.1084/jem.177.4.965.

Brodeur, G. M., A. A. Green, F. A. Hayes, K. J. Williams, D. L. Williams, and A. A. Tsiatis. 1981. “Cytogenetic Features of Human Neuroblastomas and Cell Lines.” *Cancer Research* 41 (11 Pt 1): 4678–86. https://www.ncbi.nlm.nih.gov/pubmed/6171342.

Brodeur, G. M., J. Pritchard, F. Berthold, N. L. Carlsen, V. Castel, R. P. Castelberry, B. De Bernardi, A. E. Evans, M. Favrot, and F. Hedborg. 1993. “Revisions of the International Criteria for Neuroblastoma Diagnosis, Staging, and Response to Treatment.” *Journal of Clinical Oncology*. https://doi.org/10.1200/jco.1993.11.8.1466.

Brodeur, G. M., R. C. Seeger, A. Barrett, F. Berthold, R. P. Castleberry, G. D’Angio, B. De Bernardi, A. E. Evans, M. Favrot, and A. I. Freeman. 1988. “International Criteria for Diagnosis, Staging, and Response to Treatment in Patients with Neuroblastoma.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 6 (12): 1874–81. https://doi.org/10.1200/JCO.1988.6.12.1874.

Brodeur, G. M., R. C. Seeger, M. Schwab, H. E. Varmus, and J. M. Bishop. 1984. “Amplification of N-Myc in Untreated Human Neuroblastomas Correlates with Advanced Disease Stage.” *Science* 224 (4653): 1121–24. https://doi.org/10.1126/science.6719137.

Brodeur, G. M., G. Sekhon, and M. N. Goldstein. 1977. “Chromosomal Aberrations in Human Neuroblastomas.” *Cancer* 40 (5): 2256–63. https://doi.org/3.0.co;2-1">10.1002/1097-0142(197711)40:5<2256::aid-cncr2820400536

>3.0.co;2-1.
Bronner, Marianne E., and Marcos Simões-Costa. 2016. “Chapter Seven - The Neural Crest Migrating into the Twenty-First Century.” In *Current Topics in Developmental Biology*, edited by Paul M. Wassarman, 116:115–34. Academic Press. https://doi.org/10.1016/bs.ctdb.2015.12.003.
Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering.” *American Journal of Human Genetics* 81 (5): 1084–97. https://doi.org/10.1086/521987.
Brown, P. C., S. M. Beverley, and R. T. Schimke. 1981. “Relationship of Amplified Dihydrofolate Reductase Genes to Double Minute Chromosomes in Unstably Resistant Mouse Fibroblast Cell Lines.” *Molecular and Cellular Biology* 1 (12): 1077–83. https://doi.org/10.1128/mcb.1.12.1077.
Bruin, Elza C. de, Nicholas McGranahan, Richard Mitter, Max Salm, David C. Wedge, Lucy Yates, Mariam Jamal-Hanjani, et al. 2014. “Spatial and Temporal Diversity in Genomic Instability Processes Defines Lung Cancer Evolution.” *Science* 346 (6206): 251–56. https://doi.org/10.1126/science.1253462.
Bryan, T. M., A. Englezou, J. Gupta, S. Bacchetti, and R. R. Reddel. 1995. “Telomere Elongation in Immortal Human Cells without Detectable Telomerase Activity.” *The EMBO Journal* 14 (17): 4240–48. https://doi.org/10.1002/j.1460-2075.1995.tb00098.x.
Bryan, Tracy M., Anna Englezou, Luciano Dalla-Pozza, Melissa A. Dunham, and Roger R. Reddel. 1997. “Evidence for an Alternative Mechanism for Maintaining Telomere Length in Human Tumors and Tumor-Derived Cell Lines.” *Nature Medicine* 3 (11): 1271–74. https://doi.org/10.1038/nm1197-1271.
Buccitelli, Christopher, and Matthias Selbach. 2020. “mRNAs, Proteins and the Emerging Principles of Gene Expression Control.” *Nature Reviews. Genetics*, July. https://doi.org/10.1038/s41576-020-0258-4.
Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.
Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 109 (January): 21.29.1–21.29.9. https://doi.org/10.1002/0471142727.mb2129s109.
Bunting, Samuel F., and Andre Nussenzweig. 2013. “End-Joining, Translocations and Cancer.” *Nature Reviews. Cancer* 13 (7): 443–54. https://doi.org/10.1038/nrc3537.
Cai, Yanyan, Jonathan Crowther, Tibor Pastor, Layka Abbasi Asbagh, Maria Francesca Baietti, Magdalena De Troyer, Iria Vazquez, et al. 2016. “Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism.” *Cancer Cell* 29 (5): 751–66. https://doi.org/10.1016/j.ccell.2016.04.003.
Cameron, Daniel L., Leon Di Stefano, and Anthony T. Papenfuss. 2019. “Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software.” *Nature Communications* 10 (1): 3240. https://doi.org/10.1038/s41467-019-11146-4.
Campino, Susana, Julian Forton, Srilakshmi Raj, Bert Mohr, Sarah Auburn, Andrew Fry, Valentina D. Mangano, et al. 2008. “Validating Discovered Cis-Acting Regulatory Genetic Variants: Application of an Allele Specific Expression Approach to HapMap Populations.” *PloS One* 3 (12): e4105. https://doi.org/10.1371/journal.pone.0004105.
Capasso, Mario, Marcella Devoto, Cuiping Hou, Shahab Asgharzadeh, Joseph T. Glessner, Edward F. Attiyeh, Yael P. Mosse, et al. 2009. “Common Variations in BARD1 Influence

Susceptibility to High-Risk Neuroblastoma." *Nature Genetics* 41 (6): 718–23. https://doi.org/10.1038/ng.374.
Capasso, Mario, Sharon J. Diskin, Francesca Totaro, Luca Longo, Marilena De Mariano, Roberta Russo, Flora Cimmino, et al. 2013. "Replication of GWAS-Identified Neuroblastoma Risk Loci Strengthens the Role of BARD1 and Affirms the Cumulative Effect of Genetic Variations on Disease Susceptibility." *Carcinogenesis* 34 (3): 605–11. https://doi.org/10.1093/carcin/bgs380.
Carroll, S. M., M. L. DeRose, P. Gaudray, C. M. Moore, D. R. Needham-Vandevanter, D. D. Von Hoff, and G. M. Wahl. 1988. "Double Minute Chromosomes Can Be Produced from Precursors Derived from a Chromosomal Deletion." *Molecular and Cellular Biology* 8 (4): 1525–33. https://doi.org/10.1128/mcb.8.4.1525.
Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. "Tools and Best Practices for Data Processing in Allelic Expression Analysis." *Genome Biology* 16 (September): 195. https://doi.org/10.1186/s13059-015-0762-6.
Castel, V., P. García-Miguel, A. Cañete, C. Melero, A. Navajas, J. I. Ruíz-Jiménez, S. Navarro, and M. D. Badal. 1999. "Prospective Evaluation of the International Neuroblastoma Staging System (INSS) and the International Neuroblastoma Response Criteria (INRC) in a Multicentre Setting." *European Journal of Cancer* 35 (4): 606–11. https://doi.org/10.1016/s0959-8049(98)00395-5.
Cates, C. A., R. L. Michael, K. R. Stayrook, K. A. Harvey, Y. D. Burke, S. K. Randall, P. L. Crowell, and D. N. Crowell. 1996. "Prenylation of Oncogenic Human PTP(CAAX) Protein Tyrosine Phosphatases." *Cancer Letters* 110 (1-2): 49–55. https://doi.org/10.1016/s0304-3835(96)04459-x.
Chang, Xiao, Yan Zhao, Cuiping Hou, Joseph Glessner, Lee McDaniel, Maura A. Diamond, Kelly Thomas, et al. 2017. "Common Variants in MMP20 at 11q22.2 Predispose to 11q Deletion and Neuroblastoma Risk." *Nature Communications* 8 (1): 569. https://doi.org/10.1038/s41467-017-00408-8.
Cheng, J. M., J. L. Hiemstra, S. S. Schneider, A. Naumova, N. K. Cheung, S. L. Cohn, L. Diller, C. Sapienza, and G. M. Brodeur. 1993. "Preferential Amplification of the Paternal Allele of the N-Myc Gene in Human Neuroblastomas." *Nature Genetics* 4 (2): 191–94. https://doi.org/10.1038/ng0693-191.
Cheng, Lei, Ping Wang, Sheng Yang, Yanqing Yang, Qing Zhang, Wen Zhang, Huasheng Xiao, Hengjun Gao, and Qinghua Zhang. 2012. "Identification of Genes with a Correlation between Copy Number and Expression in Gastric Cancer." *BMC Medical Genomics* 5 (May): 14. https://doi.org/10.1186/1755-8794-5-14.
Chen, Guangbo, Wahid A. Mulla, Andrei Kucharavy, Hung-Ji Tsai, Boris Rubinstein, Juliana Conkright, Scott McCroskey, et al. 2015. "Targeting the Adaptability of Heterogeneous Aneuploids." *Cell* 160 (4): 771–84. https://doi.org/10.1016/j.cell.2015.01.026.
Chen, Kevin, Erik van Nimwegen, Nikolaus Rajewsky, and Mark L. Siegal. 2010. "Correlating Gene Expression Variation with Cis-Regulatory Polymorphism in Saccharomyces Cerevisiae." *Genome Biology and Evolution* 2 (September): 697–707. https://doi.org/10.1093/gbe/evq054.
Chen, Liying, Gabriela Alexe, Neekesh V. Dharia, Linda Ross, Amanda Balboni Iniguez, Amy Saur Conway, Emily Jue Wang, et al. 2018. "CRISPR-Cas9 Screen Reveals a MYCN-Amplified Neuroblastoma Dependency on EZH2." *The Journal of Clinical Investigation* 128 (1): 446–62. https://doi.org/10.1172/JCI90793.
Chen, Lu, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, et al. 2016. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells." *Cell* 167 (5): 1398–1414.e24. https://doi.org/10.1016/j.cell.2016.10.026.

Chen, Ping, Jicheng Zhao, Yan Wang, Min Wang, Haizhen Long, Dan Liang, Li Huang, et al. 2013. “H3.3 Actively Marks Enhancers and Primes Gene Transcription via Opening Higher-Ordered Chromatin.” *Genes & Development* 27 (19): 2109–24. https://doi.org/10.1101/gad.222174.113.

Chen, Qing-Rong, Sven Bilke, Jun S. Wei, Craig C. Whiteford, Nicola Cenacchi, Alexei L. Krasnoselsky, Braden T. Greer, et al. 2004. “cDNA Array-CGH Profiling Identifies Genomic Alterations Specific to Stage and MYCN-Amplification in Neuroblastoma.” *BMC Genomics* 5 (September): 70. https://doi.org/10.1186/1471-2164-5-70.

Chen, Xiao-Jun, Hong Zhang, Zhi-Ping Tan, Wen Hu, and Yi-Feng Yang. 2016. “Novel Mutation of EXT2 Identified in a Large Family with Multiple Osteochondromas.” *Molecular Medicine Reports* 14 (5): 4687–91. https://doi.org/10.3892/mmr.2016.5814.

Chen, Yuyan, Junko Takita, Young Lim Choi, Motohiro Kato, Miki Ohira, Masashi Sanada, Lili Wang, et al. 2008. “Oncogenic Mutations of ALK Kinase in Neuroblastoma.” *Nature* 455 (7215): 971–74. https://doi.org/10.1038/nature07399.

Cheung, Nai-Kong V., Jinghui Zhang, Charles Lu, Matthew Parker, Armita Bahrami, Satish K. Tickoo, Adriana Heguy, et al. 2012. “Association of Age at Diagnosis and Genetic Mutations in Patients with Neuroblastoma.” *JAMA: The Journal of the American Medical Association* 307 (10): 1062–71. https://doi.org/10.1001/jama.2012.228.

Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. “The Impact of Structural Variation on Human Gene Expression.” *Nature Genetics* 49 (5): 692–99. https://doi.org/10.1038/ng.3834.

Chilukamarri, Laxmi, Anne L. Hancock, Sally Malik, Joanna Zabkiewicz, Jenny A. Baker, Alexander Greenhough, Anthony R. Dallosso, et al. 2007. “Hypomethylation and Aberrant Expression of the Glioma Pathogenesis-Related 1 Gene in Wilms Tumors.” *Neoplasia* 9 (11): 970–78. https://doi.org/10.1593/neo.07661.

Chipumuro, Edmond, Eugenio Marco, Camilla L. Christensen, Nicholas Kwiatkowski, Tinghu Zhang, Clark M. Hatheway, Brian J. Abraham, et al. 2014. “CDK7 Inhibition Suppresses Super-Enhancer-Linked Oncogenic Transcription in MYCN-Driven Cancer.” *Cell* 159 (5): 1126–39. https://doi.org/10.1016/j.cell.2014.10.024.

Chong, Zechen, Jue Ruan, Min Gao, Wanding Zhou, Tenghui Chen, Xian Fan, Li Ding, et al. 2017. “novoBreak: Local Assembly for Breakpoint Detection in Cancer Genomes.” *Nature Methods* 14 (1): 65–67. https://doi.org/10.1038/nmeth.4084.

Choy, Mun-Kit, Mehregan Movassagh, Hock-Guan Goh, Martin R. Bennett, Thomas A. Down, and Roger S. Y. Foo. 2010. “Genome-Wide Conserved Consensus Transcription Factor Binding Motifs Are Hyper-Methylated.” *BMC Genomics* 11 (September): 519. https://doi.org/10.1186/1471-2164-11-519.

Chunduri, Narendra Kumar, and Zuzana Storchová. 2019. “The Diverse Consequences of Aneuploidy.” *Nature Cell Biology* 21 (1): 54–62. https://doi.org/10.1038/s41556-018-0243-8.

Ciriello, Giovanni, Martin L. Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. 2013. “Emerging Landscape of Oncogenic Signatures across Human Cancers.” *Nature Genetics* 45 (10): 1127–33. https://doi.org/10.1038/ng.2762.

Clynes, David, Douglas R. Higgs, and Richard J. Gibbons. 2013. “The Chromatin Remodeller ATRX: A Repeat Offender in Human Disease.” *Trends in Biochemical Sciences* 38 (9): 461–66. https://doi.org/10.1016/j.tibs.2013.06.011.

Clynes, David, Clare Jelinska, Barbara Xella, Helena Ayyub, Caroline Scott, Matthew Mitson, Stephen Taylor, Douglas R. Higgs, and Richard J. Gibbons. 2015. “Suppression of the Alternative Lengthening of Telomere Pathway by the Chromatin Remodelling Factor ATRX.” *Nature Communications* 6 (July): 7538. https://doi.org/10.1038/ncomms8538.

Cohen, Sarit, Neta Agmon, Olga Sobol, and Daniel Segal. 2010. “Extrachromosomal Circles of Satellite Repeats and 5S Ribosomal DNA in Human Cells.” *Mobile DNA* 1 (1): 11. https://doi.org/10.1186/1759-8753-1-11.

Cohen, Sarit, Andreas Houben, and Daniel Segal. 2007. “Extrachromosomal Circular DNA Derived from Tandemly Repeated Genomic Sequences in Plants: Extrachromosomal DNA Circles in Plants.” *The Plant Journal: For Cell and Molecular Biology* 53 (6): 1027–34. https://doi.org/10.1111/j.1365-313X.2007.03394.x.

Contrepois, Kévin, Clément Coudereau, Bérénice A. Benayoun, Nadine Schuler, Pierre-François Roux, Oliver Bischof, Régis Courbeyrette, et al. 2017. “Histone Variant H2A.J Accumulates in Senescent Cells and Promotes Inflammatory Gene Expression.” *Nature Communications* 8 (May): 14995. https://doi.org/10.1038/ncomms14995.

Coquelle, Arnaud, Lorène Rozier, Bernard Dutrillaux, and Michelle Debatisse. 2002. “Induction of Multiple Double-Strand Breaks within an Hsr by meganucleaseI-SceI Expression or Fragile Site Activation Leads to Formation of Double Minutes and Other Chromosomal Rearrangements.” *Oncogene* 21 (50): 7671–79. https://doi.org/10.1038/sj.onc.1205880.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.

Corradin, Olivia, and Peter C. Scacheri. 2014. “Enhancer Variants: Evaluating Functions in Common Disease.” *Genome Medicine* 6 (10): 85. https://doi.org/10.1186/s13073-014-0085-3.

Cortés-Ciriano, Isidro, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. “Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing.” *Nature Genetics* 52 (3): 331–41. https://doi.org/10.1038/s41588-019-0576-7.

Corvi, R., L. Savelyeva, L. Amler, R. Handgretinger, and M. Schwab. 1995. “Cytogenetic Evolution of MYCN and MDM2 Amplification in the Neuroblastoma LS Tumour and Its Cell Line.” *European Journal of Cancer* 31A (4): 520–23. https://doi.org/10.1016/0959-8049(95)00031-d.

Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert. 1978. “Molecular Basis of Base Substitution Hotspots in Escherichia Coli.” *Nature* 274 (5673): 775–80. https://doi.org/10.1038/274775a0.

Cox, D. R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 34 (2): 187–220. http://www.jstor.org/stable/2985181.

Cox, D., C. Yuncken, and A. I. Spriggs. 1965. “MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD.” *The Lancet* 1 (7402): 55–58. https://doi.org/10.1016/s0140-6736(65)90131-5.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. “Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21931–36. https://doi.org/10.1073/pnas.1016071107.

Curwen, Val, Eduardo Eyras, T. Daniel Andrews, Laura Clarke, Emmanuel Mongin, Steven M. J. Searle, and Michele Clamp. 2004. “The Ensembl Automatic Gene Annotation System.” *Genome Research* 14 (5): 942–50. https://doi.org/10.1101/gr.1858004.

Cuzzoni, E., L. Ferretti, C. Giordani, S. Castiglione, and F. Sala. 1990. “A Repeated Chromosomal DNA Sequence Is Amplified as a Circular Extrachromosomal Molecule in Rice (Oryza Sativa L.).” *Molecular & General Genetics: MGG* 222 (1): 58–64. https://doi.org/10.1007/BF00283023.

Dai, Juncheng, Meng Zhu, Cheng Wang, Wei Shen, Wen Zhou, Jie Sun, Jia Liu, et al. 2015. “Systematical Analyses of Variants in CTCF-Binding Sites Identified a Novel Lung Cancer Susceptibility Locus among Chinese Population.” *Scientific Reports* 5 (January): 7833. https://doi.org/10.1038/srep07833.

Dalla-Favera, R., M. Bregni, J. Erikson, D. Patterson, R. C. Gallo, and C. M. Croce. 1982. “Human c-Myc Onc Gene Is Located on the Region of Chromosome 8 That Is Translocated in Burkitt Lymphoma Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (24): 7824–27. https://doi.org/10.1073/pnas.79.24.7824.

Davis, Erica, Florian Caiment, Xavier Tordoir, Jérôme Cavaillé, Anne Ferguson-Smith, Noelle Cockett, Michel Georges, and Carole Charlier. 2005. “RNAi-Mediated Allelic Trans-Interaction at the Imprinted Rtl1/Peg11 Locus.” *Current Biology: CB* 15 (8): 743–49. https://doi.org/10.1016/j.cub.2005.02.060.

Davoli, Teresa, Andrew Wei Xu, Kristen E. Mengwasser, Laura M. Sack, John C. Yoon, Peter J. Park, and Stephen J. Elledge. 2013. “Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome.” *Cell* 155 (4): 948–62. https://doi.org/10.1016/j.cell.2013.10.011.

Deaton, Aimée M., and Adrian Bird. 2011. “CpG Islands and the Regulation of Transcription.” *Genes & Development* 25 (10): 1010–22. https://doi.org/10.1101/gad.2037511.

Decaesteker, Bieke, Geertrui Denecker, Christophe Van Neste, Emmy M. Dolman, Wouter Van Loocke, Moritz Gartlgruber, Carolina Nunes, et al. 2018. “TBX2 Is a Neuroblastoma Core Regulatory Circuitry Component Enhancing MYCN/FOXM1 Reactivation of DREAM Targets.” *Nature Communications* 9 (1): 4866. https://doi.org/10.1038/s41467-018-06699-9.

deCarvalho, Ana C., Hoon Kim, Laila M. Poisson, Mary E. Winn, Claudius Mueller, David Cherba, Julie Koeman, et al. 2018. “Discordant Inheritance of Chromosomal and Extrachromosomal DNA Elements Contributes to Dynamic Disease Evolution in Glioblastoma.” *Nature Genetics* 50 (5): 708–17. https://doi.org/10.1038/s41588-018-0105-0.

Decock, Anneleen, Maté Ongenaert, Jo Vandesompele, and Frank Speleman. 2011. “Neuroblastoma Epigenetics: From Candidate Gene Approaches to Genome-Wide Screenings.” *Epigenetics: Official Journal of the DNA Methylation Society* 6 (8): 962–70. https://doi.org/10.4161/epi.6.8.16516.

Degner, Jacob F., John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. 2009. “Effect of Read-Mapping Biases on Detecting Allele-Specific Expression from RNA-Sequencing Data.” *Bioinformatics* 25 (24): 3207–12. https://doi.org/10.1093/bioinformatics/btp579.

Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. “DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation.” *Nature* 482 (7385): 390–94. https://doi.org/10.1038/nature10808.

Demars, Julie, Mansur Ennuri Shmela, Sylvie Rossignol, Jun Okabe, Irène Netchine, Salah Azzi, Sylvie Cabrol, et al. 2010. “Analysis of the IGF2/H19 Imprinting Control Region Uncovers New Genetic Defects, Including Mutations of OCT-Binding Sequences, in Patients with 11p15 Fetal Growth Disorders.” *Human Molecular Genetics* 19 (5): 803–14. https://doi.org/10.1093/hmg/ddp549.

Dermitzakis, Emmanouil T., and Barbara E. Stranger. 2006. “Genetic Variation in Human Gene Expression.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 17 (6): 503–8. https://doi.org/10.1007/s00335-006-0005-y.

Dilley, Robert L., Priyanka Verma, Nam Woo Cho, Harrison D. Winters, Anne R. Wondisford, and Roger A. Greenberg. 2016. “Break-Induced Telomere Synthesis Underlies

Alternative Telomere Maintenance." *Nature* 539 (7627): 54–58. https://doi.org/10.1038/nature20099.
Dillon, Laura W., Pankaj Kumar, Yoshiyuki Shibata, Yuh-Hwa Wang, Smaranda Willcox, Jack D. Griffith, Yves Pommier, Shunichi Takeda, and Anindya Dutta. 2015. "Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity." *Cell Reports* 11 (11): 1749–59. https://doi.org/10.1016/j.celrep.2015.05.020.
Ding, Zhihao, Massimo Mangino, Abraham Aviv, Tim Spector, Richard Durbin, and UK10K Consortium. 2014. "Estimating Telomere Length from Whole Genome Sequence Data." *Nucleic Acids Research* 42 (9): e75. https://doi.org/10.1093/nar/gku181.
Ding, Zhihao, Yunyun Ni, Sander W. Timmer, Bum-Kyu Lee, Anna Battenhouse, Sandra Louzada, Fengtang Yang, et al. 2014. "Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association." *PLoS Genetics* 10 (11): e1004798. https://doi.org/10.1371/journal.pgen.1004798.
Diskin, Sharon J., Mario Capasso, Maura Diamond, Derek A. Oldridge, Karina Conkrite, Kristopher R. Bosse, Mike R. Russell, et al. 2014. "Rare Variants in TP53 and Susceptibility to Neuroblastoma." *Journal of the National Cancer Institute* 106 (4): dju047. https://doi.org/10.1093/jnci/dju047.
Diskin, Sharon J., Mario Capasso, Robert W. Schnepp, Kristina A. Cole, Edward F. Attiyeh, Cuiping Hou, Maura Diamond, et al. 2012. "Common Variation at 6q16 within HACE1 and LIN28B Influences Susceptibility to Neuroblastoma." *Nature Genetics* 44 (10): 1126–30. https://doi.org/10.1038/ng.2387.
Diskin, Sharon J., Cuiping Hou, Joseph T. Glessner, Edward F. Attiyeh, Marci Laudenslager, Kristopher Bosse, Kristina Cole, et al. 2009. "Copy Number Variation at 1q21.1 Associated with Neuroblastoma." *Nature* 459 (7249): 987–91. https://doi.org/10.1038/nature08035.
Dixon, Anna L., Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny C. C. Wong, Jenny Taylor, et al. 2007. "A Genome-Wide Association Study of Global Gene Expression." *Nature Genetics* 39 (10): 1202–7. https://doi.org/10.1038/ng2109.
Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.
Dodt, Matthias, Johannes T. Roehr, Rina Ahmed, and Christoph Dieterich. 2012. "FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms." *Biology* 1 (3): 895–905. https://doi.org/10.3390/biology1030895.
Drier, Yotam, Matthew J. Cotton, Kaylyn E. Williamson, Shawn M. Gillespie, Russell J. H. Ryan, Michael J. Kluk, Christopher D. Carey, et al. 2016. "An Oncogenic MYB Feedback Loop Drives Alternate Cell Fates in Adenoid Cystic Carcinoma." *Nature Genetics* 48 (3): 265–72. https://doi.org/10.1038/ng.3502.
Durbin, Adam D., Mark W. Zimmerman, Neekesh V. Dharia, Brian J. Abraham, Amanda Balboni Iniguez, Nina Weichert-Leahey, Shuning He, et al. 2018. "Selective Gene Dependencies in MYCN-Amplified Neuroblastoma Include the Core Transcriptional Regulatory Circuitry." *Nature Genetics* 50 (9): 1240–46. https://doi.org/10.1038/s41588-018-0191-z.
Edfors, Fredrik, Frida Danielsson, Björn M. Hallström, Lukas Käll, Emma Lundberg, Fredrik Pontén, Björn Forsström, and Mathias Uhlén. 2016. "Gene-Specific Correlation of RNA and Protein Levels in Human Cells and Tissues." *Molecular Systems Biology* 12 (10): 883. https://doi.org/10.15252/msb.20167144.
Emigh, T. H. 1980. "A Comparison of Tests for Hardy-Weinberg Equilibrium." *Biometrics* 36 (4): 627–42. https://doi.org/10.2307/2556115.

Emilsson, Valur, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, et al. 2008. “Genetics of Gene Expression and Its Effect on Disease.” *Nature* 452 (7186): 423–28. https://doi.org/10.1038/nature06758.

ENCODE Project Consortium. 2011. “A User’s Guide to the Encyclopedia of DNA Elements (ENCODE).” *PLoS Biology* 9 (4): e1001046. https://doi.org/10.1371/journal.pbio.1001046.

———. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods* 9 (3): 215–16. https://doi.org/10.1038/nmeth.1906.

Evans, Audrey E., Jane Chatten, Giulio J. D’Angio, James M. Gerson, Janis Robinson, and Louise Schnaufer. 1980. “A Review of 17 IV-S Neuroblastoma Patients at the Children’s Hospital of Philadelphia.” *Cancer* 45 (5): 833–39. https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142(19800301)45:5%3C833::AID-CNCR2820450502%3E3.0.CO;2-U.

Fairfax, Benjamin P., Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O. Vannberg, and Julian C. Knight. 2012. “Genetics of Gene Expression in Primary Immune Cells Identifies Cell Type-Specific Master Regulators and Roles of HLA Alleles.” *Nature Genetics* 44 (5): 502–10. https://doi.org/10.1038/ng.2205.

Fan, Biao, Somkid Dachrut, Ho Coral, Siu Tsan Yuen, Kent Man Chu, Simon Law, Lianhai Zhang, Jiafu Ji, Suet Yi Leung, and Xin Chen. 2012. “Integration of DNA Copy Number Alterations and Transcriptional Expression Analysis in Human Gastric Cancer.” *PloS One* 7 (4): e29824. https://doi.org/10.1371/journal.pone.0029824.

Fang, Hai-Tong, Chadi A. El Farran, Qiao Rui Xing, Li-Feng Zhang, Hu Li, Bing Lim, and Yuin-Han Loh. 2018. “Global H3.3 Dynamic Deposition Defines Its Bimodal Role in Cell Fate Transition.” *Nature Communications* 9 (1): 1537. https://doi.org/10.1038/s41467-018-03904-7.

Fan, Guobiao, Dan Ye, Songcheng Zhu, Jiajie Xi, Xudong Guo, Jing Qiao, Yukang Wu, et al. 2017. “RTL1 Promotes Melanoma Proliferation by Regulating Wnt/β-Catenin Signalling.” *Oncotarget* 8 (62): 106026–37. https://doi.org/10.18632/oncotarget.22523.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. “Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants.” *Nature* 518 (7539): 337–43. https://doi.org/10.1038/nature13835.

Faust, Gregory G., and Ira M. Hall. 2014. “SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction.” *Bioinformatics* 30 (17): 2503–5. https://doi.org/10.1093/bioinformatics/btu314.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. 2015. “Sequenza: Allele-Specific Copy Number and Mutation Profiles from Tumor Sequencing Data.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (1): 64–70. https://doi.org/10.1093/annonc/mdu479.

Fehrmann, Rudolf S. N., Juha M. Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H. Pers, et al. 2015. “Gene Expression Analysis Identifies Global Gene Dosage Sensitivity in Cancer.” *Nature Genetics* 47 (2): 115–25. https://doi.org/10.1038/ng.3173.

Fellermann, Klaus, Daniel E. Stange, Elke Schaeffeler, Hartmut Schmalzl, Jan Wehkamp, Charles L. Bevins, Walter Reinisch, et al. 2006. “A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon.” *American Journal of Human Genetics* 79 (3): 439–48.

https://doi.org/10.1086/505915.
Feng, Jianxing, Tao Liu, and Yong Zhang. 2011. “Using MACS to Identify Peaks from ChIP-Seq Data.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 2 (June): Unit 2.14. https://doi.org/10.1002/0471250953.bi0214s34.
Ferrari, Francesco, Artyom A. Alekseyenko, Peter J. Park, and Mitzi I. Kuroda. 2014. “Transcriptional Control of a Whole Chromosome: Emerging Models for Dosage Compensation.” *Nature Structural & Molecular Biology* 21 (2): 118–25. https://doi.org/10.1038/nsmb.2763.
Flavahan, William A., Yotam Drier, Brian B. Liau, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvà, and Bradley E. Bernstein. 2016. “Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas.” *Nature* 529 (7584): 110–14. https://doi.org/10.1038/nature16490.
Fletez-Brant, Christopher, Dongwon Lee, Andrew S. McCallion, and Michael A. Beer. 2013. “Kmer-SVM: A Web Server for Identifying Predictive Regulatory Sequence Features in Genomic Data Sets.” *Nucleic Acids Research* 41 (Web Server issue): W544–56. https://doi.org/10.1093/nar/gkt519.
Flores, S. C., T. K. Moore, and J. W. Gaubatz. 1987. “Dispersed Repetitive Sequences of the Mouse Genome Are Differentially Represented in Extrachromosomal Circular DNAs in Vivo.” *Plasmid* 17 (3): 257–60. https://doi.org/10.1016/0147-619x(87)90034-5.
Fogarty, Marie P., Rui Xiao, Ludmila Prokunina-Olsson, Laura J. Scott, and Karen L. Mohlke. 2010. “Allelic Expression Imbalance at High-Density Lipoprotein Cholesterol Locus MMAB-MVK.” *Human Molecular Genetics* 19 (10): 1921–29. https://doi.org/10.1093/hmg/ddq067.
Fong, C. T., N. C. Dracopoli, P. S. White, P. T. Merrill, R. C. Griffith, D. E. Housman, and G. M. Brodeur. 1989. “Loss of Heterozygosity for the Short Arm of Chromosome 1 in Human Neuroblastomas: Correlation with N-Myc Amplification.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (10): 3753–57. https://doi.org/10.1073/pnas.86.10.3753.
Fortelny, Nikolaus, Christopher M. Overall, Paul Pavlidis, and Gabriela V. Cohen Freue. 2017. “Can We Predict Protein from mRNA Levels?” *Nature* 547 (7664): E19–20. https://doi.org/10.1038/nature22293.
Frank, Derk, Detlef Doenecke, and Werner Albig. 2003. “Differential Expression of Human Replacement and Cell Cycle Dependent H3 Histone Genes.” *Gene* 312 (July): 135–43. https://doi.org/10.1016/s0378-1119(03)00609-7.
Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. “Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications.” *Nature* 538 (7624): 265–69. https://doi.org/10.1038/nature19800.
Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73. https://doi.org/10.1093/nar/gky955.
Fredriksson, Nils J., Lars Ny, Jonas A. Nilsson, and Erik Larsson. 2014. “Systematic Analysis of Noncoding Somatic Mutations and Gene Expression Alterations across 14 Tumor Types.” *Nature Genetics* 46 (12): 1258–63. https://doi.org/10.1038/ng.3141.
Freedman, Matthew L., David Reich, Kathryn L. Penney, Gavin J. McDonald, Andre A. Mignault, Nick Patterson, Stacey B. Gabriel, et al. 2004. “Assessing the Impact of Population Stratification on Genetic Association Studies.” *Nature Genetics* 36 (4): 388–93. https://doi.org/10.1038/ng1333.
Fröhlich, Leopold F., Maria Mrakovcic, Ralf Steinborn, Ung-Il Chung, Murat Bastepe, and

Harald Jüppner. 2010. “Targeted Deletion of the Nesp55 DMR Defines Another Gnas Imprinting Control Region and Provides a Mouse Model of Autosomal Dominant PHP-Ib.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9275–80. https://doi.org/10.1073/pnas.0910224107.

Fujita, Toshitsugu, Junko Ikuta, Juri Hamada, Toshihide Okajima, Kenji Tatematsu, Katsuyuki Tanizawa, and Shun ’ichi Kuroda. 2004. “Identification of a Tissue-Non-Specific Homologue of Axonal Fasciculation and Elongation Protein Zeta-1.” *Biochemical and Biophysical Research Communications* 313 (3): 738–44. https://doi.org/10.1016/j.bbrc.2003.12.006.

Fullwood, Melissa J., Chia-Lin Wei, Edison T. Liu, and Yijun Ruan. 2009. “Next-Generation DNA Sequencing of Paired-End Tags (PET) for Transcriptome and Genome Analyses.” *Genome Research* 19 (4): 521–32. https://doi.org/10.1101/gr.074906.107.

Furlan, Alessandro, Moritz Lübke, Igor Adameyko, Francois Lallemend, and Patrik Ernfors. 2013. “The Transcription Factor Hmx1 and Growth Factor Receptor Activities Control Sympathetic Neurons Diversification.” *The EMBO Journal* 32 (11): 1613–25. https://doi.org/10.1038/emboj.2013.85.

Gamazon, Eric R., Ayellet V. Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S. Xi, Farhad Hormozdiari, Halit Ongen, et al. 2018. “Using an Atlas of Gene Regulation across 44 Human Tissues to Inform Complex Disease- and Trait-Associated Variation.” *Nature Genetics* 50 (7): 956–67. https://doi.org/10.1038/s41588-018-0154-4.

Gaubatz, J. W., and S. C. Flores. 1990. “Tissue-Specific and Age-Related Variations in Repetitive Sequences of Mouse Extrachromosomal Circular DNAs.” *Mutation Research* 237 (1): 29–36. https://doi.org/10.1016/0921-8734(90)90029-q.

Ge, Bing, Dmitry K. Pokholok, Tony Kwan, Elin Grundberg, Lisanne Morcos, Dominique J. Verlaan, Jennie Le, et al. 2009. “Global Patterns of Cis Variation in Human Cells Revealed by High-Density Allelic Expression Analysis.” *Nature Genetics* 41 (11): 1216–22. https://doi.org/10.1038/ng.473.

George, Rani E., Takaomi Sanda, Megan Hanna, Stefan Fröhling, William Luther 2nd, Jianming Zhang, Yebin Ahn, et al. 2008. “Activating Mutations in ALK Provide a Therapeutic Target in Neuroblastoma.” *Nature* 455 (7215): 975–78. https://doi.org/10.1038/nature07397.

George, S. L., V. Parmar, F. Lorenzi, L. V. Marshall, Y. Jamin, E. Poon, P. Angelini, and L. Chesler. 2020. “Novel Therapeutic Strategies Targeting Telomere Maintenance Mechanisms in High-Risk Neuroblastoma.” *Journal of Experimental & Clinical Cancer Research: CR* 39 (1): 78. https://doi.org/10.1186/s13046-020-01582-2.

Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, M. Math, James Larkin, David Endesfelder, Eva Gronroos, et al. 2012. “Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing.” *The New England Journal of Medicine* 366 (10): 883–92. https://doi.org/10.1056/NEJMoa1113205.

Ghandi, Mahmoud, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. 2014. “Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features.” *PLoS Computational Biology* 10 (7): e1003711. https://doi.org/10.1371/journal.pcbi.1003711.

Giam, Maybelline, and Giulia Rancati. 2015. “Aneuploidy and Chromosomal Instability in Cancer: A Jackpot to Chaos.” *Cell Division* 10 (May): 3. https://doi.org/10.1186/s13008-015-0009-7.

Gilad, Yoav, Scott A. Rifkin, and Jonathan K. Pritchard. 2008. “Revealing the Architecture of Gene Regulation: The Promise of eQTL Studies.” *Trends in Genetics: TIG* 24 (8): 408–15. https://doi.org/10.1016/j.tig.2008.06.001.

Gilbert, F., M. Feder, G. Balaban, D. Brangman, D. K. Lurie, R. Podolsky, V. Rinaldt, N. Vinikoor, and J. Weisband. 1984. “Human Neuroblastomas and Abnormalities of Chromosomes 1 and 17.” *Cancer Research* 44 (11): 5444–49.

https://www.ncbi.nlm.nih.gov/pubmed/6488196.
Gil, Noa, and Igor Ulitsky. 2020. “Regulation of Gene Expression by Cis-Acting Long Non-Coding RNAs.” *Nature Reviews. Genetics* 21 (2): 102–17. https://doi.org/10.1038/s41576-019-0184-5.
Giorgio, Elisa, Daniel Robyr, Malte Spielmann, Enza Ferrero, Eleonora Di Gregorio, Daniele Imperiale, Giovanna Vaula, et al. 2015. “A Large Genomic Deletion Leads to Enhancer Adoption by the Lamin B1 Gene: A Second Path to Autosomal Dominant Adult-Onset Demyelinating Leukodystrophy (ADLD).” *Human Molecular Genetics* 24 (11): 3143–54. https://doi.org/10.1093/hmg/ddv065.
Giresi, Paul G., Jonghwan Kim, Ryan M. McDaniell, Vishwanath R. Iyer, and Jason D. Lieb. 2007. “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin.” *Genome Research* 17 (6): 877–85. https://doi.org/10.1101/gr.5533506.
Giwa, Abdulazeez, Azeez Fatai, Junaid Gamieldien, Alan Christoffels, and Hocine Bendou. 2020. “Identification of Novel Prognostic Markers of Survival Time in High-Risk Neuroblastoma Using Gene Expression Profiles.” *Oncotarget* 11 (46): 4293–4305. https://doi.org/10.18632/oncotarget.27808.
Goldberg, Aaron D., Laura A. Banaszynski, Kyung-Min Noh, Peter W. Lewis, Simon J. Elsaesser, Sonja Stadler, Scott Dewell, et al. 2010. “Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions.” *Cell* 140 (5): 678–91. https://doi.org/10.1016/j.cell.2010.01.003.
Gonzalez, Enrique, Hemant Kulkarni, Hector Bolivar, Andrea Mangano, Racquel Sanchez, Gabriel Catano, Robert J. Nibbs, et al. 2005. “The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility.” *Science* 307 (5714): 1434–40. https://doi.org/10.1126/science.1101160.
Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51. https://doi.org/10.1038/nrg.2016.49.
Göring, Harald H. H., Joanne E. Curran, Matthew P. Johnson, Thomas D. Dyer, Jac Charlesworth, Shelley A. Cole, Jeremy B. M. Jowett, et al. 2007. “Discovery of Expression QTLs Using Large-Scale Transcriptional Profiling in Human Lymphocytes.” *Nature Genetics* 39 (10): 1208–16. https://doi.org/10.1038/ng2119.
Graf, Marco, Diego Bonetti, Arianna Lockhart, Kamar Serhal, Vanessa Kellner, André Maicher, Pascale Jolivet, Maria Teresa Teixeira, and Brian Luke. 2017. “Telomere Length Determines TERRA and R-Loop Regulation through the Cell Cycle.” *Cell* 170 (1): 72–85.e14. https://doi.org/10.1016/j.cell.2017.06.006.
Grassi, Elisa S., Pauline Jeannot, Vasiliki Pantazopoulou, Tracy J. Berg, and Alexander Pietras. 2020. “Niche-Derived Soluble DLK1 Promotes Glioma Growth.” *Neoplasia* 22 (12): 689–701. https://doi.org/10.1016/j.neo.2020.10.005.
Greenman, Chris D. 2012. “Cancer. Haploinsufficient Gene Selection in Cancer.” *Science*. https://doi.org/10.1126/science.1224806.
Groningen, Tim van, Jan Koster, Linda J. Valentijn, Danny A. Zwijnenburg, Nurdan Akogul, Nancy E. Hasselt, Marloes Broekmans, et al. 2017. “Neuroblastoma Is Composed of Two Super-Enhancer-Associated Differentiation States.” *Nature Genetics* 49 (8): 1261–66. https://doi.org/10.1038/ng.3899.
Grubert, Fabian, Judith B. Zaugg, Maya Kasowski, Oana Ursu, Damek V. Spacek, Alicia R. Martin, Peyton Greenside, et al. 2015. “Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions.” *Cell* 162 (5): 1051–65. https://doi.org/10.1016/j.cell.2015.07.048.
Grundberg, Elin, Kerrin S. Small, Åsa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, et al. 2012. “Mapping Cis- and Trans-Regulatory Effects

across Multiple Tissues in Twins." *Nature Genetics* 44 (10): 1084–89. https://doi.org/10.1038/ng.2394.
Gryder, Berkley E., Marco Wachtel, Kenneth Chang, Osama El Demerdash, Nicholas G. Aboreden, Wardah Mohammed, Winston Ewert, et al. 2020. "Miswired Enhancer Logic Drives a Cancer of the Muscle Lineage." *iScience* 23 (5): 101103. https://doi.org/10.1016/j.isci.2020.101103.
Guenther, Catherine A., Bosiljka Tasic, Liqun Luo, Mary A. Bedell, and David M. Kingsley. 2014. "A Molecular Basis for Classic Blond Hair Color in Europeans." *Nature Genetics* 46 (7): 748–52. https://doi.org/10.1038/ng.2991.
Guo, Yu Amanda, Mei Mei Chang, Weitai Huang, Wen Fong Ooi, Manjie Xing, Patrick Tan, and Anders Jacobsen Skanderup. 2018. "Mutation Hotspots at CTCF Binding Sites Coupled to Chromosomal Instability in Gastrointestinal Cancers." *Nature Communications* 9 (1): 1520. https://doi.org/10.1038/s41467-018-03828-2.
Haeussler, Maximilian, Ann S. Zweig, Cath Tyner, Matthew L. Speir, Kate R. Rosenbloom, Brian J. Raney, Christopher M. Lee, et al. 2019. "The UCSC Genome Browser Database: 2019 Update." *Nucleic Acids Research* 47 (D1): D853–58. https://doi.org/10.1093/nar/gky1095.
Hahn, W. C., C. M. Counter, A. S. Lundberg, R. L. Beijersbergen, M. W. Brooks, and R. A. Weinberg. 1999. "Creation of Human Tumour Cells with Defined Genetic Elements." *Nature* 400 (6743): 464–68. https://doi.org/10.1038/22780.
Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.
Hanahan, D., and R. A. Weinberg. 2000. "The Hallmarks of Cancer." *Cell* 100 (1): 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.
Hannon, Eilis, Tyler J. Gorrie-Stone, Melissa C. Smart, Joe Burrage, Amanda Hughes, Yanchun Bao, Meena Kumari, Leonard C. Schalkwyk, and Jonathan Mill. 2018. "Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits." *American Journal of Human Genetics* 103 (5): 654–65. https://doi.org/10.1016/j.ajhg.2018.09.007.
Hannon, Eilis, Mike Weedon, Nicholas Bray, Michael O'Donovan, and Jonathan Mill. 2017. "Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci." *American Journal of Human Genetics* 100 (6): 954–59. https://doi.org/10.1016/j.ajhg.2017.04.013.
Hark, A. T., C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman. 2000. "CTCF Mediates Methylation-Sensitive Enhancer-Blocking Activity at the H19/Igf2 Locus." *Nature* 405 (6785): 486–89. https://doi.org/10.1038/35013106.
Harrow, Jennifer, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, et al. 2006. "GENCODE: Producing a Reference Annotation for ENCODE." *Genome Biology* 7 Suppl 1 (August): S4.1–9. https://doi.org/10.1186/gb-2006-7-s1-s4.
Hartlieb, Sabine A., Lina Sieverling, Michal Nadler-Holly, Matthias Ziehm, Umut H. Toprak, Carl Herrmann, Naveed Ishaque, et al. 2021. "Alternative Lengthening of Telomeres in Childhood Neuroblastoma from Genome to Proteome." *Nature Communications* 12 (1): 1269. https://doi.org/10.1038/s41467-021-21247-8.
Hasin-Brumshtein, Yehudit, Farhad Hormozdiari, Lisa Martin, Atila van Nas, Eleazar Eskin, Aldons J. Lusis, and Thomas A. Drake. 2014. "Allele-Specific Expression and eQTL Analysis in Mouse Adipose Tissue." *BMC Genomics* 15 (June): 471. https://doi.org/10.1186/1471-2164-15-471.
Hasty, Paul, and Cristina Montagna. 2014. "Chromosomal Rearrangements in Cancer: Detection and Potential Causal Mechanisms." *Molecular & Cellular Oncology* 1 (1). https://doi.org/10.4161/mco.29904.

Heaphy, Christopher M., Roeland F. de Wilde, Yuchen Jiao, Alison P. Klein, Barish H. Edil, Chanjuan Shi, Chetan Bettegowda, et al. 2011. “Altered Telomeres in Tumors with ATRX and DAXX Mutations.” *Science* 333 (6041): 425. https://doi.org/10.1126/science.1207313.

Hecht, J. T., D. Hogue, Y. Wang, S. H. Blanton, M. Wagner, L. C. Strong, W. Raskind, M. F. Hansen, and D. Wells. 1997. “Hereditary Multiple Exostoses (EXT): Mutational Studies of Familial EXT1 Cases and EXT-Associated Malignancies.” *American Journal of Human Genetics* 60 (1): 80–86. https://www.ncbi.nlm.nih.gov/pubmed/8981950.

Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. “Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome.” *Nature Genetics* 39 (3): 311–18. https://doi.org/10.1038/ng1966.

Heinz, Sven, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. 2015. “The Selection and Function of Cell Type-Specific Enhancers.” *Nature Reviews. Molecular Cell Biology* 16 (3): 144–54. https://doi.org/10.1038/nrm3949.

Helmsauer, Konstantin, Maria E. Valieva, Salaheddine Ali, Rocío Chamorro González, Robert Schöpflin, Claudia Röefzaad, Yi Bei, et al. 2020. “Enhancer Hijacking Determines Extrachromosomal Circular MYCN Amplicon Architecture in Neuroblastoma.” *Nature Communications* 11 (1): 5823. https://doi.org/10.1038/s41467-020-19452-y.

Henckel, Amandine, Kazuhiko Nakabayashi, Lionel A. Sanz, Robert Feil, Kenichiro Hata, and Philippe Arnaud. 2009. “Histone Methylation Is Mechanistically Linked to DNA Methylation at Imprinting Control Regions in Mammals.” *Human Molecular Genetics* 18 (18): 3375–83. https://doi.org/10.1093/hmg/ddp277.

Henrich, Kai-Oliver, Sebastian Bender, Maral Saadati, Daniel Dreidax, Moritz Gartlgruber, Chunxuan Shao, Carl Herrmann, et al. 2016. “Integrative Genome-Scale Analysis Identifies Epigenetic Mechanisms of Transcriptional Deregulation in Unfavorable Neuroblastomas.” *Cancer Research* 76 (18): 5523–37. https://doi.org/10.1158/0008-5472.CAN-15-2507.

Henrichsen, Charlotte N., Evelyne Chaignat, and Alexandre Reymond. 2009. “Copy Number Variants, Diseases and Gene Expression.” *Human Molecular Genetics* 18 (R1): R1–8. https://doi.org/10.1093/hmg/ddp011.

Henrichsen, Charlotte N., Nicolas Vinckenbosch, Sebastian Zöllner, Evelyne Chaignat, Sylvain Pradervand, Frédéric Schütz, Manuel Ruedi, Henrik Kaessmann, and Alexandre Reymond. 2009. “Segmental Copy Number Variation Shapes Tissue Transcriptomes.” *Nature Genetics* 41 (4): 424–29. https://doi.org/10.1038/ng.345.

Henson, Jeremy D., Ying Cao, Lily I. Huschtscha, Andy C. Chang, Amy Y. M. Au, Hilda A. Pickett, and Roger R. Reddel. 2009. “DNA C-Circles Are Specific and Quantifiable Markers of Alternative-Lengthening-of-Telomeres Activity.” *Nature Biotechnology* 27 (12): 1181–85. https://doi.org/10.1038/nbt.1587.

Hertwig, Falk, Martin Peifer, and Matthias Fischer. 2016. “Telomere Maintenance Is Pivotal for High-Risk Neuroblastoma.” *Cell Cycle* 15 (3): 311–12. https://doi.org/10.1080/15384101.2015.1125243.

He, Xiaoyu, Shanyu Chen, Ruilin Li, Xinyin Han, Zhipeng He, Danyang Yuan, Shuying Zhang, Xiaohong Duan, and Beifang Niu. 2020. “Comprehensive Fundamental Somatic Variant Calling and Quality Management Strategies for Human Cancer Genomes.” *Briefings in Bioinformatics*, June. https://doi.org/10.1093/bib/bbaa083.

Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106

(23): 9362–67. https://doi.org/10.1073/pnas.0903103106.
Hiyama, Eiso, Keiko Hiyama, Takashi Yokoyama, Yuichiro Matsuura, Mieczyslaw A. Piatyszek, and Jerry W. Shay. 1995. “Correlating Telomerase Activity Levels with Human Neuroblastoma Outcomes.” *Nature Medicine*. https://doi.org/10.1038/nm0395-249.
Hnisz, Denes, Abraham S. Weintraub, Daniel S. Day, Anne-Laure Valton, Rasmus O. Bak, Charles H. Li, Johanna Goldmann, et al. 2016. “Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods.” *Science* 351 (6280): 1454–58. https://doi.org/10.1126/science.aad9024.
Hong, Xiaohong, Na Liu, Yelin Liang, Qingmei He, Xiaojing Yang, Yuan Lei, Panpan Zhang, et al. 2020. “Circular RNA CRIM1 Functions as a ceRNA to Promote Nasopharyngeal Carcinoma Metastasis and Docetaxel Chemoresistance through Upregulating FOXQ1.” *Molecular Cancer* 19 (1): 33. https://doi.org/10.1186/s12943-020-01149-x.
Horlings, Hugo M., Carmen Lai, Dimitry S. A. Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A. Joosse, et al. 2010. “Integration of DNA Copy Number Alterations and Prognostic Gene Expression Signatures in Breast Cancer Patients.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 16 (2): 651–63. https://doi.org/10.1158/1078-0432.CCR-09-0709.
Horn, Susanne, Adina Figl, P. Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, et al. 2013. “TERT Promoter Mutations in Familial and Sporadic Melanoma.” *Science* 339 (6122): 959–61. https://doi.org/10.1126/science.1230062.
Horowitz, H., and J. E. Haber. 1985. “Identification of Autonomously Replicating Circular Subtelomeric Y’ Elements in Saccharomyces Cerevisiae.” *Molecular and Cellular Biology* 5 (9): 2369–80. https://doi.org/10.1128/mcb.5.9.2369.
Huang, Franklin W., Eran Hodis, Mary Jue Xu, Gregory V. Kryukov, Lynda Chin, and Levi A. Garraway. 2013. “Highly Recurrent TERT Promoter Mutations in Human Melanoma.” *Science* 339 (6122): 957–59. https://doi.org/10.1126/science.1229259.
Huang, F. W., C. M. Bielski, M. L. Rinne, W. C. Hahn, W. R. Sellers, F. Stegmeier, L. A. Garraway, and G. V. Kryukov. 2015. “TERT Promoter Mutations and Monoallelic Activation of TERT in Cancer.” *Oncogenesis* 4 (December): e176. https://doi.org/10.1038/oncsis.2015.39.
Hubertus, Jochen, Martin Lacher, Marietta Rottenkolber, Josef Müller-Höcker, Michael Berger, Maximilian Stehr, Dietrich von Schweinitz, and Roland Kappler. 2011. “Altered Expression of Imprinted Genes in Wilms Tumors.” *Oncology Reports* 25 (3): 817–23. https://doi.org/10.3892/or.2010.1113.
Hui, R., D. H. Campbell, C. S. Lee, K. McCaul, D. J. Horsfall, E. A. Musgrove, R. J. Daly, R. Seshadri, and R. L. Sutherland. 1997. “EMS1 Amplification Can Occur Independently of CCND1 or INT-2 Amplification at 11q13 and May Identify Different Phenotypes in Primary Breast Cancer.” *Oncogene* 15 (13): 1617–23. https://doi.org/10.1038/sj.onc.1201311.
Hyun, Kwangbeom, Jongcheol Jeon, Kihyun Park, and Jaehoon Kim. 2017. “Writing, Erasing and Reading Histone Lysine Methylations.” *Experimental & Molecular Medicine* 49 (4): e324. https://doi.org/10.1038/emm.2017.11.
Iafrate, A. John, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. 2004. “Detection of Large-Scale Variation in the Human Genome.” *Nature Genetics* 36 (9): 949–51. https://doi.org/10.1038/ng1416.
Ibrahim, Mahmoud M., Scott A. Lacadie, and Uwe Ohler. 2015. “JAMM: A Peak Finder for Joint Analysis of NGS Replicates.” *Bioinformatics* 31 (1): 48–55. https://doi.org/10.1093/bioinformatics/btu568.
Imataka, George, and Osamu Arisaka. 2012. “Chromosome Analysis Using Spectral

Karyotyping (SKY)." *Cell Biochemistry and Biophysics* 62 (1): 13–17. https://doi.org/10.1007/s12013-011-9285-2.
International Human Genome Sequencing Consortium. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.
Jakubosky, David, Matteo D'Antonio, Marc Jan Bonder, Craig Smail, Margaret K. R. Donovan, William W. Young Greenwald, Hiroko Matsui, et al. 2020. "Properties of Structural Variants and Short Tandem Repeats Associated with Gene Expression and Complex Traits." *Nature Communications* 11 (1): 2927. https://doi.org/10.1038/s41467-020-16482-4.
Jamal-Hanjani, Mariam, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B. K. Watkins, Selvaraju Veeriah, Seema Shafi, et al. 2017. "Tracking the Evolution of Non-Small-Cell Lung Cancer." *The New England Journal of Medicine* 376 (22): 2109–21. https://doi.org/10.1056/NEJMoa1616288.
Janoueix-Lerosey, Isabelle, Gudrun Schleiermacher, Evi Michels, Véronique Mosseri, Agnès Ribeiro, Delphine Lequin, Joëlle Vermeulen, et al. 2009. "Overall Genomic Pattern Is a Predictor of Outcome in Neuroblastoma." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (7): 1026–33. https://doi.org/10.1200/JCO.2008.16.0630.
Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 316 (5830): 1497–1502. https://doi.org/10.1126/science.1141319.
Jones, R. S., and S. S. Potter. 1985. "L1 Sequences in HeLa Extrachromosomal Circular DNA: Evidence for Circularization by Homologous Recombination." *Proceedings of the National Academy of Sciences of the United States of America* 82 (7): 1989–93. https://doi.org/10.1073/pnas.82.7.1989.
Junnila, Siina, Arto Kokkola, Marja-Liisa Karjalainen-Lindsberg, Pauli Puolakkainen, and Outi Monni. 2010. "Genome-Wide Gene Copy Number and Expression Analysis of Primary Gastric Tumors and Gastric Cancer Cell Lines." *BMC Cancer* 10 (March): 73. https://doi.org/10.1186/1471-2407-10-73.
Kaczówka, Przemysław, Aleksandra Wieczorek, Małgorzata Czogała, Teofila Książek, Katarzyna Szewczyk, and Walentyna Balwierz. 2018. "The Role of N-Myc Gene Amplification in Neuroblastoma Childhood Tumour - Single-Centre Experience." *Contemporary Oncology* 22 (4): 223–28. https://doi.org/10.5114/wo.2018.81402.
Karabacak Calviello, Aslıhan, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. 2019. "Reproducible Inference of Transcription Factor Footprints in ATAC-Seq and DNase-Seq Datasets Using Protocol-Specific Bias Modeling." *Genome Biology* 20 (1): 42. https://doi.org/10.1186/s13059-019-1654-y.
Karlseder, J., R. Zeillinger, C. Schneeberger, K. Czerwenka, P. Speiser, E. Kubista, D. Birnbaum, P. Gaudray, and C. Theillet. 1994. "Patterns of DNA Amplification at Band q13 of Chromosome 11 in Human Breast Cancer." *Genes, Chromosomes & Cancer* 9 (1): 42–48. https://doi.org/10.1002/gcc.2870090108.
Karlsson, Jenny, Anders Valind, Linda Holmquist Mengelbier, Sofia Bredin, Louise Cornmark, Caroline Jansson, Amina Wali, et al. 2018. "Four Evolutionary Trajectories Underlie Genetic Intratumoral Variation in Childhood Cancer." *Nature Genetics* 50 (7): 944–50. https://doi.org/10.1038/s41588-018-0131-y.
Katainen, Riku, Kashyap Dave, Esa Pitkänen, Kimmo Palin, Teemu Kivioja, Niko Välimäki, Alexandra E. Gylfe, et al. 2015. "CTCF/cohesin-Binding Sites Are Frequently Mutated in Cancer." *Nature Genetics* 47 (7): 818–21. https://doi.org/10.1038/ng.3335.
Katzenstein, H. M., L. C. Bowman, G. M. Brodeur, P. S. Thorner, V. V. Joshi, E. I. Smith, A. T. Look, et al. 1998. "Prognostic Significance of Age, MYCN Oncogene Amplification,

Tumor Cell Ploidy, and Histology in 110 Infants with Stage D(S) Neuroblastoma: The Pediatric Oncology Group Experience--a Pediatric Oncology Group Study." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 16 (6): 2007–17. https://doi.org/10.1200/JCO.1998.16.6.2007.

Kaufman, R. J., P. C. Brown, and R. T. Schimke. 1979. "Amplified Dihydrofolate Reductase Genes in Unstably Methotrexate-Resistant Cells Are Associated with Double Minute Chromosomes." *Proceedings of the National Academy of Sciences of the United States of America* 76 (11): 5669–73. https://doi.org/10.1073/pnas.76.11.5669.

Kelley, David R., Jasper Snoek, and John L. Rinn. 2016. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks." *Genome Research* 26 (7): 990–99. https://doi.org/10.1101/gr.200535.115.

Ke, Xiao-Xue, Dunke Zhang, Hailong Zhao, Renjian Hu, Zhen Dong, Rui Yang, Shunqin Zhu, Qingyou Xia, Han-Fei Ding, and Hongjuan Cui. 2015. "Phox2B Correlates with MYCN and Is a Prognostic Marker for Neuroblastoma Development." *Oncology Letters* 9 (6): 2507–14. https://doi.org/10.3892/ol.2015.3088.

Killela, Patrick J., Zachary J. Reitman, Yuchen Jiao, Chetan Bettegowda, Nishant Agrawal, Luis A. Diaz Jr, Allan H. Friedman, et al. 2013. "TERT Promoter Mutations Occur Frequently in Gliomas and a Subset of Tumors Derived from Cells with Low Rates of Self-Renewal." *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): 6021–26. https://doi.org/10.1073/pnas.1303607110.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." *Nature Methods* 12 (4): 357–60. https://doi.org/10.1038/nmeth.3317.

Kim, N. W., M. A. Piatyszek, K. R. Prowse, C. B. Harley, M. D. West, P. L. Ho, G. M. Coviello, W. E. Wright, S. L. Weinrich, and J. W. Shay. 1994. "Specific Association of Human Telomerase Activity with Immortal Cells and Cancer." *Science* 266 (5193): 2011–15. https://doi.org/10.1126/science.7605428.

Kim, Tae-Min, Seung-Hyun Jung, Chang Hyeok An, Sung Hak Lee, In-Pyo Baek, Min Sung Kim, Sung-Won Park, Je-Keun Rhee, Sug-Hyung Lee, and Yeun-Jun Chung. 2015. "Subclonal Genomic Architectures of Primary and Metastatic Colorectal Cancer Based on Intratumoral Genetic Heterogeneity." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21 (19): 4461–72. https://doi.org/10.1158/1078-0432.CCR-14-2413.

Kinoshita, Yasuhiro, Noboru Ohnishi, Yasuyuki Yamada, Takahiro Kunisada, and Hideo Yamagishi. 1985. "Extrachromosomal Circular DNA from Nuclear Fraction of Higher Plants." *Plant & Cell Physiology* 26 (7): 1401–9. https://doi.org/10.1093/oxfordjournals.pcp.a077040.

Kitazawa, Moe, Akito Sutani, Tomoko Kaneko-Ishino, and Fumitoshi Ishino. 2021. "The Role of Eutherian-Specific RTL1 in the Nervous System and Its Implications for the Kagami-Ogata and Temple Syndromes." *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, January. https://doi.org/10.1111/gtc.12830.

Kittles, Rick A., Weidong Chen, Ramesh K. Panguluri, Chiledum Ahaghotu, Aaron Jackson, Clement A. Adebamowo, Robin Griffin, et al. 2002. "CYP3A4-V and Prostate Cancer in African Americans: Causal or Confounding Association because of Population Stratification?" *Human Genetics* 110 (6): 553–60. https://doi.org/10.1007/s00439-002-0731-5.

Klambauer, Günter, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. 2012. "cn.MOPS: Mixture of Poissons for Discovering Copy Number Variations in next-Generation Sequencing Data with a Low False Discovery Rate." *Nucleic Acids Research* 40 (9): e69. https://doi.org/10.1093/nar/gks003.

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. 2012. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research* 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111.

Koche, Richard P., Elias Rodriguez-Fos, Konstantin Helmsauer, Martin Burkert, Ian C. MacArthur, Jesper Maag, Rocio Chamorro, et al. 2020. “Extrachromosomal Circular DNA Drives Oncogenic Genome Remodeling in Neuroblastoma.” *Nature Genetics* 52 (1): 29–34. https://doi.org/10.1038/s41588-019-0547-z.

Kohl, N. E., N. Kanda, R. R. Schreck, G. Bruns, S. A. Latt, F. Gilbert, and F. W. Alt. 1983. “Transposition and Amplification of Oncogene-Related Sequences in Human Neuroblastomas.” *Cell* 35 (2 Pt 1): 359–67. https://doi.org/10.1016/0092-8674(83)90169-1.

Kolle, G., K. Georgas, G. P. Holmes, M. H. Little, and T. Yamada. 2000. “CRIM1, a Novel Gene Encoding a Cysteine-Rich Repeat Protein, Is Developmentally Regulated and Implicated in Vertebrate CNS Development and Organogenesis.” *Mechanisms of Development* 90 (2): 181–93. https://doi.org/10.1016/s0925-4773(99)00248-8.

Koneru, Balakrishna, Gonzalo Lopez, Ahsan Farooqi, Karina L. Conkrite, Thinh H. Nguyen, Shawn J. Macha, Apexa Modi, et al. 2020. “Telomere Maintenance Mechanisms Define Clinical Outcome in High-Risk Neuroblastoma.” *Cancer Research* 80 (12): 2663–75. https://doi.org/10.1158/0008-5472.CAN-19-3068.

Kong, Jinhwa, Jaemoon Shin, Jungim Won, Keonbae Lee, Unjoo Lee, and Jeehee Yoon. 2017. “ExCNVSS: A Noise-Robust Method for Copy Number Variation Detection in Whole Exome Sequencing Data.” *BioMed Research International* 2017 (June): 9631282. https://doi.org/10.1155/2017/9631282.

Korbel, Jan O., and Peter J. Campbell. 2013. “Criteria for Inference of Chromothripsis in Cancer Genomes.” *Cell* 152 (6): 1226–36. https://doi.org/10.1016/j.cell.2013.02.023.

Kornberg, Roger D. 2007. “The Molecular Basis of Eukaryotic Transcription.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (32): 12955–61. https://doi.org/10.1073/pnas.0704138104.

Korotkevich, Gennady, Vladimir Sukhov, and Alexey Sergushichev. 2019. “Fast Gene Set Enrichment Analysis.” *bioRxiv*. https://doi.org/10.1101/060012.

Köster, Johannes, and Sven Rahmann. 2012. “Snakemake--a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Krolewski, J. J., and M. G. Rush. 1984. “Some Extrachromosomal Circular DNAs Containing the Alu Family of Dispersed Repetitive Sequences May Be Reverse Transcripts.” *Journal of Molecular Biology* 174 (1): 31–40. https://doi.org/10.1016/0022-2836(84)90363-2.

Kunisada, T., and H. Yamagishi. 1984. “Sequence Repetition and Genomic Distribution of Small Polydisperse Circular DNA Purified from HeLa Cells.” *Gene* 31 (1-3): 213–23. https://doi.org/10.1016/0378-1119(84)90212-9.

———. 1987. “Sequence Organization of Repetitive Sequences Enriched in Small Polydisperse Circular DNAs from HeLa Cells.” *Journal of Molecular Biology* 198 (4): 557–65. https://doi.org/10.1016/0022-2836(87)90199-9.

Kunisada, T., H. Yamagishi, Z. Ogita, T. Kirakawa, and Y. Mitsui. 1985. “Appearance of Extrachromosomal Circular DNAs during in Vivo and in Vitro Ageing of Mammalian Cells.” *Mechanisms of Ageing and Development* 29 (1): 89–99. https://doi.org/10.1016/0047-6374(85)90050-8.

Laborda, J., E. A. Sausville, T. Hoffman, and V. Notario. 1993. “Dlk, a Putative Mammalian Homeotic Gene Differentially Expressed in Small Cell Lung Carcinoma and

Neuroendocrine Tumor Cell Line." *The Journal of Biological Chemistry* 268 (6): 3817–20. https://www.ncbi.nlm.nih.gov/pubmed/8095043.
Lachner, M., D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. 2001. "Methylation of Histone H3 Lysine 9 Creates a Binding Site for HP1 Proteins." *Nature* 410 (6824): 116–20. https://doi.org/10.1038/35065132.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. https://doi.org/10.1186/gb-2009-10-3-r25.
Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11. https://doi.org/10.1038/nature12531.
Łastowska, M., V. Viprey, M. Santibanez-Koref, I. Wappler, H. Peters, C. Cullinane, P. Roberts, et al. 2007. "Identification of Candidate Genes Involved in Neuroblastoma Progression by Combining Genomic and Expression Microarrays with Survival Data." *Oncogene* 26 (53): 7432–44. https://doi.org/10.1038/sj.onc.1210552.
Latos, Paulina A., Stefan H. Stricker, Laura Steenpass, Florian M. Pauler, Ru Huang, Basak H. Senergin, Kakkad Regha, et al. 2009. "An in Vitro ES Cell Imprinting Model Shows That Imprinted Expression of the Igf2r Gene Arises from an Allele-Specific Expression Bias." *Development* 136 (3): 437–48. https://doi.org/10.1242/dev.032060.
Lauberth, Shannon M., Takahiro Nakayama, Xiaolin Wu, Andrea L. Ferris, Zhanyun Tang, Stephen H. Hughes, and Robert G. Roeder. 2013. "H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation." *Cell* 152 (5): 1021–36. https://doi.org/10.1016/j.cell.2013.01.052.
Lau, D. T., L. B. Hesson, M. D. Norris, G. M. Marshall, M. Haber, and L. J. Ashton. 2012. "Prognostic Significance of Promoter DNA Methylation in Patients with Childhood Neuroblastoma." *Clinical Cancer Research*. https://doi.org/10.1158/1078-0432.ccr-12-0294.
Lawrence, Michael, Robert Gentleman, and Vincent Carey. 2009. "Rtracklayer: An R Package for Interfacing with Genome Browsers." *Bioinformatics* 25 (14): 1841–42. https://doi.org/10.1093/bioinformatics/btp328.
Lawrenson, Kate, Siddhartha Kar, Karen McCue, Karoline Kuchenbaeker, Kyriaki Michailidou, Jonathan Tyrer, Jonathan Beesley, et al. 2016. "Functional Mechanisms Underlying Pleiotropic Risk Alleles at the 19p13.1 Breast-Ovarian Cancer Susceptibility Locus." *Nature Communications* 7 (September): 12675. https://doi.org/10.1038/ncomms12675.
Lawrenson, Kate, Qiyuan Li, Siddhartha Kar, Ji-Heui Seo, Jonathan Tyrer, Tassja J. Spindler, Janet Lee, et al. 2015. "Cis-eQTL Analysis and Functional Validation of Candidate Susceptibility Genes for High-Grade Serous Ovarian Cancer." *Nature Communications* 6 (September): 8234. https://doi.org/10.1038/ncomms9234.
Lee, Bum-Kyu, and Vishwanath R. Iyer. 2012. "Genome-Wide Studies of CCCTC-Binding Factor (CTCF) and Cohesin Provide Insight into Chromatin Structure and Regulation." *The Journal of Biological Chemistry* 287 (37): 30906–13. https://doi.org/10.1074/jbc.R111.324962.
Lee, D. Y., J. J. Hayes, D. Pruss, and A. P. Wolffe. 1993. "A Positive Role for Histone Acetylation in Transcription Factor Access to Nucleosomal DNA." *Cell* 72 (1): 73–84. https://doi.org/10.1016/0092-8674(93)90051-q.
Leeuwen, F. N., H. E. Kain, R. A. Kammen, F. Michiels, O. W. Kranenburg, and J. G. Collard. 1997. "The Guanine Nucleotide Exchange Factor Tiam1 Affects Neuronal Morphology; Opposing Roles for the Small GTPases Rac and Rho." *The Journal of Cell Biology* 139 (3): 797–807. https://doi.org/10.1083/jcb.139.3.797.

Lee, W. H., A. L. Murphree, and W. F. Benedict. 1984. “Expression and Amplification of the N-Myc Gene in Primary Retinoblastoma.” *Nature* 309 (5967): 458–60. https://doi.org/10.1038/309458a0.

Lewis, Peter W., Simon J. Elsaesser, Kyung-Min Noh, Sonja C. Stadler, and C. David Allis. 2010. “Daxx Is an H3.3-Specific Histone Chaperone and Cooperates with ATRX in Replication-Independent Chromatin Assembly at Telomeres.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (32): 14075–80. https://doi.org/10.1073/pnas.1008850107.

Liang, Gangning, Joy C. Y. Lin, Vivian Wei, Christine Yoo, Jonathan C. Cheng, Carvell T. Nguyen, Daniel J. Weisenberger, et al. 2004. “Distinct Localization of Histone H3 Acetylation and H3-K4 Methylation to the Transcription Start Sites in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7357–62. https://doi.org/10.1073/pnas.0401866101.

Libermann, T. A., H. R. Nusbaum, N. Razon, R. Kris, I. Lax, H. Soreq, N. Whittle, M. D. Waterfield, A. Ullrich, and J. Schlessinger. 1985. “Amplification, Enhanced Expression and Possible Rearrangement of EGF Receptor Gene in Primary Human Brain Tumours of Glial Origin.” *Nature* 313 (5998): 144–47. https://doi.org/10.1038/313144a0.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Jason, Richard Lupat, Kaushalya C. Amarasinghe, Ella R. Thompson, Maria A. Doyle, Georgina L. Ryland, Richard W. Tothill, Saman K. Halgamuge, Ian G. Campbell, and Kylie L. Gorringe. 2012. “CONTRA: Copy Number Analysis for Targeted Resequencing.” *Bioinformatics* 28 (10): 1307–13. https://doi.org/10.1093/bioinformatics/bts146.

Lindblad-Toh, K., D. M. Tanenbaum, M. J. Daly, E. Winchester, W. O. Lui, A. Villapakkam, S. E. Stanton, et al. 2000. “Loss-of-Heterozygosity Analysis of Small-Cell Lung Carcinomas Using Single-Nucleotide Polymorphism Arrays.” *Nature Biotechnology* 18 (9): 1001–5. https://doi.org/10.1038/79269.

Lindeboom, Rik G. H., Fran Supek, and Ben Lehner. 2016. “The Rules and Impact of Nonsense-Mediated mRNA Decay in Human Cancers.” *Nature Genetics* 48 (10): 1112–18. https://doi.org/10.1038/ng.3664.

Lippert, C., F. P. Casale, B. Rakitsch, and O. Stegle. 2014. “LIMIX: Genetic Analysis of Multiple Traits.” *bioRxiv*. https://doi.org/10.1101/003905.

Lippert, Christoph, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. 2011. “FaST Linear Mixed Models for Genome-Wide Association Studies.” *Nature Methods* 8 (10): 833–35. https://doi.org/10.1038/nmeth.1681.

Li, Qiyuan, Ji-Heui Seo, Barbara Stranger, Aaron McKenna, Itsik Pe’er, Thomas Laframboise, Myles Brown, Svitlana Tyekucheva, and Matthew L. Freedman. 2013. “Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci.” *Cell* 152 (3): 633–41. https://doi.org/10.1016/j.cell.2012.12.034.

Li, Qiyuan, Alexander Stram, Constance Chen, Siddhartha Kar, Simon Gayther, Paul Pharoah, Christopher Haiman, Barbara Stranger, Peter Kraft, and Matthew L. Freedman. 2014. “Expression QTL-Based Analyses Reveal Candidate Causal Genes and Loci across Five Tumor Types.” *Human Molecular Genetics* 23 (19): 5294–5302. https://doi.org/10.1093/hmg/ddu228.

Liu, Eric Minwei, Alexander Martinez-Fundichely, Bianca Jay Diaz, Boaz Aronson, Tawny Cuykendall, Matthew MacKay, Priyanka Dhingra, et al. 2019. “Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes.” *Cell Systems* 8 (5): 446–55.e8. https://doi.org/10.1016/j.cels.2019.04.001.

Liu, Yu, Chong Chen, Zhengmin Xu, Claudio Scuoppo, Cory D. Rillahan, Jianjiong Gao, Barbara Spitzer, et al. 2016. “Deletions Linked to TP53 Loss Drive Cancer through

p53-Independent Mechanisms." *Nature* 531 (7595): 471–75. https://doi.org/10.1038/nature17157.
Liu, Zhi, Jing Yang, Huayong Xu, Chao Li, Zhen Wang, Yuanyuan Li, Xiao Dong, and Yixue Li. 2014. "Comparing Computational Methods for Identification of Allele-Specific Expression Based on next Generation Sequencing Data." *Genetic Epidemiology* 38 (7): 591–98. https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21846.
Li, Yilong, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, et al. 2020. "Patterns of Somatic Structural Variation in Human Cancer Genomes." *Nature* 578 (7793): 112–21. https://doi.org/10.1038/s41586-019-1913-9.
Li, Z., B. J. Abraham, A. Berezovskaya, N. Farah, Y. Liu, T. Leon, A. Fielding, et al. 2017. "APOBEC Signature Mutation Generates an Oncogenic Enhancer That Drives LMO1 Expression in T-ALL." *Leukemia* 31 (10): 2057–64. https://doi.org/10.1038/leu.2017.75.
Lloyd-Jones, Luke R., Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, et al. 2017. "The Genetic Architecture of Gene Expression in Peripheral Blood." *American Journal of Human Genetics* 100 (2): 228–37. https://doi.org/10.1016/j.ajhg.2016.12.008.
Loh, Po-Ru, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, et al. 2016. "Reference-Based Phasing Using the Haplotype Reference Consortium Panel." *Nature Genetics* 48 (11): 1443–48. https://doi.org/10.1038/ng.3679.
López-Maury, Luis, Samuel Marguerat, and Jürg Bähler. 2008. "Tuning Gene Expression to Changing Environments: From Rapid Responses to Evolutionary Adaptation." *Nature Reviews. Genetics* 9 (8): 583–93. https://doi.org/10.1038/nrg2398.
Lopez-Pajares, V., A. Rubin, B. Barajas, M. Furlan-Magaril, M. Mumbach, W. Greenleaf, A. Kundaje, et al. 2017. "464 Dynamic and Stable Enhancer-Promoter Contacts Regulate Epidermal Terminal Differentiation." *Journal of Investigative Dermatology*. https://doi.org/10.1016/j.jid.2017.02.483.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.
Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." *Cell* 161 (5): 1012–25. https://doi.org/10.1016/j.cell.2015.04.004.
Ly, Peter, and Don W. Cleveland. 2017. "Rebuilding Chromosomes After Catastrophe: Emerging Mechanisms of Chromothripsis." *Trends in Cell Biology* 27 (12): 917–30. https://doi.org/10.1016/j.tcb.2017.08.005.
MacArthur, Daniel G., Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science* 335 (6070): 823–28. https://doi.org/10.1126/science.1215040.
Mac, S. M., C. A. D'Cunha, and P. J. Farnham. 2000. "Direct Recruitment of N-Myc to Target Gene Promoters." *Molecular Carcinogenesis* 29 (2): 76–86. https://doi.org/3.0.co;2-y">10.1002/1098-2744(200010)29:2<76::aid-mc4>3.0.co;2-y.
Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1): 246. https://doi.org/10.1186/s13059-019-1828-7.
Main, Bradley J., Ryan D. Bickel, Lauren M. McIntyre, Rita M. Graze, Peter P. Calabrese, and Sergey V. Nuzhdin. 2009. "Allele-Specific Expression Assays Using Solexa." *BMC Genomics* 10 (September): 422. https://doi.org/10.1186/1471-2164-10-422.

Mainieri, Avantika, and David Haig. 2019. “Retrotransposon Gag-like 1 (RTL1) and the Molecular Evolution of Self-Targeting Imprinted microRNAs.” *Biology Direct* 14 (1): 18. https://doi.org/10.1186/s13062-019-0250-0.

Mansour, Marc R., Brian J. Abraham, Lars Anders, Alla Berezovskaya, Alejandro Gutierrez, Adam D. Durbin, Julia Etchin, et al. 2014. “Oncogene Regulation. An Oncogenic Super-Enhancer Formed through Somatic Mutation of a Noncoding Intergenic Element.” *Science* 346 (6215): 1373–77. https://doi.org/10.1126/science.1259037.

Maquat, L. E. 1995. “When Cells Stop Making Sense: Effects of Nonsense Codons on RNA Metabolism in Vertebrate Cells.” *RNA* 1 (5): 453–65. https://www.ncbi.nlm.nih.gov/pubmed/7489507.

Maris, J. M., C. Guo, P. S. White, M. D. Hogarty, P. M. Thompson, D. O. Stram, R. Gerbing, K. K. Matthay, R. C. Seeger, and G. M. Brodeur. 2001. “Allelic Deletion at Chromosome Bands 11q14-23 Is Common in Neuroblastoma.” *Medical and Pediatric Oncology* 36 (1): 24–27. https://doi.org/10.1002/1096-911X(20010101)36:1<24::AID-MPO1007>3.0.CO;2-7.

Maris, J. M., P. S. White, C. P. Beltinger, E. P. Sulman, R. P. Castleberry, J. J. Shuster, A. T. Look, and G. M. Brodeur. 1995. “Significance of Chromosome 1p Loss of Heterozygosity in Neuroblastoma.” *Cancer Research* 55 (20): 4664–69. https://www.ncbi.nlm.nih.gov/pubmed/7553646.

Maris, John M., Yael P. Mosse, Jonathan P. Bradfield, Cuiping Hou, Stefano Monni, Richard H. Scott, Shahab Asgharzadeh, et al. 2008. “Chromosome 6p22 Locus Associated with Clinically Aggressive Neuroblastoma.” *The New England Journal of Medicine* 358 (24): 2585–93. https://doi.org/10.1056/NEJMoa0708698.

Marmorstein, Ronen, and Ming-Ming Zhou. 2014. “Writers and Readers of Histone Acetylation: Structure, Mechanism, and Inhibition.” *Cold Spring Harbor Perspectives in Biology* 6 (7): a018762. https://doi.org/10.1101/cshperspect.a018762.

Matthay, Katherine K., John M. Maris, Gudrun Schleiermacher, Akira Nakagawara, Crystal L. Mackall, Lisa Diller, and William A. Weiss. 2016. “Neuroblastoma.” *Nature Reviews. Disease Primers* 2 (November): 16078. https://doi.org/10.1038/nrdp.2016.78.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” *Science* 337 (6099): 1190–95. https://doi.org/10.1126/science.1222794.

Mayrhofer, Markus, Sebastian DiLorenzo, and Anders Isaksson. 2013. “Patchwork: Allele-Specific Copy Number Analysis of Whole-Genome Sequenced Tumor Tissue.” *Genome Biology* 14 (3): R24. https://doi.org/10.1186/gb-2013-14-3-r24.

McCarroll, Steven A., Alan Huett, Petric Kuballa, Shannon D. Chilewski, Aimee Landry, Philippe Goyette, Michael C. Zody, et al. 2008. “Deletion Polymorphism Upstream of IRGM Associated with Altered IRGM Expression and Crohn’s Disease.” *Nature Genetics* 40 (9): 1107–12. https://doi.org/10.1038/ng.215.

McColgan, P., and S. J. Tabrizi. 2018. “Huntington’s Disease: A Clinical Review.” *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies* 25 (1): 24–34. https://doi.org/10.1111/ene.13413.

McDaniel, Lee D., Karina L. Conkrite, Xiao Chang, Mario Capasso, Zalman Vaksman, Derek A. Oldridge, Anna Zachariou, et al. 2017. “Common Variants Upstream of MLF1 at 3q25 and within CPZ at 4p16 Associated with Neuroblastoma.” *PLoS Genetics* 13 (5): e1006787. https://doi.org/10.1371/journal.pgen.1006787.

McGaughey, David M., Zachary E. Stine, Jimmy L. Huynh, Ryan M. Vinton, and Andrew S. McCallion. 2009. “Asymmetrical Distribution of Non-Conserved Regulatory Sequences at PHOX2B Is Reflected at the ENCODE Loci and Illuminates a Possible Genome-Wide Trend.” *BMC Genomics* 10 (January): 8. https://doi.org/10.1186/1471-2164-10-8.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Meddeb, M., G. Danglot, I. Chudoba, A. M. Vénuat, J. Bénard, H. Avet-Loiseau, B. Vasseur, et al. 1996. “Additional Copies of a 25 Mb Chromosomal Region Originating from 17q23.1-17qter Are Present in 90% of High-Grade Neuroblastomas.” *Genes, Chromosomes & Cancer* 17 (3): 156–65. https://doi.org/10.1002/(SICI)1098-2264(199611)17:3<156::AID-GCC3>3.0.CO;2-3.

Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, et al. 2012. “Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay.” *Nature Biotechnology* 30 (3): 271–77. https://doi.org/10.1038/nbt.2137.

Merla, Giuseppe, Cédric Howald, Charlotte N. Henrichsen, Robert Lyle, Carine Wyss, Marie-Thérèse Zabot, Stylianos E. Antonarakis, and Alexandre Reymond. 2006. “Submicroscopic Deletion in Patients with Williams-Beuren Syndrome Influences Expression Levels of the Nonhemizygous Flanking Genes.” *American Journal of Human Genetics* 79 (2): 332–41. https://doi.org/10.1086/506371.

Metzker, Michael L. 2010. “Sequencing Technologies - the next Generation.” *Nature Reviews. Genetics* 11 (1): 31–46. https://doi.org/10.1038/nrg2626.

Mikkelsen, Tarjei S., Manching Ku, David B. Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, et al. 2007. “Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells.” *Nature* 448 (7153): 553–60. https://doi.org/10.1038/nature06008.

Milani, Lili, Manu Gupta, Malin Andersen, Sumeer Dhar, Mårten Fryknäs, Anders Isaksson, Rolf Larsson, and Ann-Christine Syvänen. 2007. “Allelic Imbalance in Gene Expression as a Guide to Cis-Acting Regulatory Single Nucleotide Polymorphisms in Cancer Cells.” *Nucleic Acids Research* 35 (5): e34. https://doi.org/10.1093/nar/gkl1152.

Milani, Lili, Anders Lundmark, Jessica Nordlund, Anna Kiialainen, Trond Flaegstad, Gudmundur Jonmundsson, Jukka Kanerva, et al. 2009. “Allele-Specific Gene Expression Patterns in Primary Leukemic Cells Reveal Regulation of Gene Expression by CpG Site Methylation.” *Genome Research* 19 (1): 1–11. https://doi.org/10.1101/gr.083931.108.

Minasi, Simone, Caterina Baldi, Francesca Gianno, Manila Antonelli, Anna Maria Buccoliero, Torsten Pietsch, Maura Massimino, and Francesca Romana Buttarelli. 2021. “Alternative Lengthening of Telomeres in Molecular Subgroups of Paediatric High-Grade Glioma.” *Child’s Nervous System: ChNS: Official Journal of the International Society for Pediatric Neurosurgery* 37 (3): 809–18. https://doi.org/10.1007/s00381-020-04933-8.

Min, Jinrong, Abdellah Allali-Hassani, Nataliya Nady, Chao Qi, Hui Ouyang, Yongsong Liu, Farrell MacKenzie, Masoud Vedadi, and Cheryl H. Arrowsmith. 2007. “L3MBTL1 Recognition of Mono- and Dimethylated Histones.” *Nature Structural & Molecular Biology* 14 (12): 1229–30. https://doi.org/10.1038/nsmb1340.

Mittal, Vinay K., and John F. McDonald. 2017. “De Novo Assembly and Characterization of Breast Cancer Transcriptomes Identifies Large Numbers of Novel Fusion-Gene Transcripts of Potential Functional Significance.” *BMC Medical Genomics* 10 (1): 53. https://doi.org/10.1186/s12920-017-0289-7.

Miyake, Izumi, Yuko Hakomori, Azusa Shinohara, Toshie Gamou, Masaki Saito, Akihiro Iwamatsu, and Ryuichi Sakai. 2002. “Activation of Anaplastic Lymphoma Kinase Is Responsible for Hyperphosphorylation of ShcC in Neuroblastoma Cell Lines.” *Oncogene* 21 (38): 5823–34. https://doi.org/10.1038/sj.onc.1205735.

Mizuno, Akira, and Yukinori Okada. 2019. “Biological Characterization of Expression

Quantitative Trait Loci (eQTLs) Showing Tissue-Specific Opposite Directional Effects." *European Journal of Human Genetics: EJHG* 27 (11): 1745–56. https://doi.org/10.1038/s41431-019-0468-4.

Mogno, Ilaria, Jamie C. Kwasnieski, and Barak A. Cohen. 2013. "Massively Parallel Synthetic Promoter Assays Reveal the in Vivo Effects of Binding Site Variants." *Genome Research* 23 (11): 1908–15. https://doi.org/10.1101/gr.157891.113.

Molenaar, Jan J., Raquel Domingo-Fernández, Marli E. Ebus, Sven Lindner, Jan Koster, Ksenija Drabek, Pieter Mestdagh, et al. 2012. "LIN28B Induces Neuroblastoma and Enhances MYCN Levels via Let-7 Suppression." *Nature Genetics* 44 (11): 1199–1206. https://doi.org/10.1038/ng.2436.

Molenaar, Jan J., Jan Koster, Danny A. Zwijnenburg, Peter van Sluis, Linda J. Valentijn, Ida van der Ploeg, Mohamed Hamdi, et al. 2012. "Sequencing of Neuroblastoma Identifies Chromothripsis and Defects in Neuritogenesis Genes." *Nature* 483 (7391): 589–93. https://doi.org/10.1038/nature10910.

Molina, Jessica, Paulina Carmona-Mora, Jacqueline Chrast, Paola M. Krall, César P. Canales, James R. Lupski, Alexandre Reymond, and Katherina Walz. 2008. "Abnormal Social Behaviors and Altered Gene Expression Rates in a Mouse Model for Potocki-Lupski Syndrome." *Human Molecular Genetics* 17 (16): 2486–95. https://doi.org/10.1093/hmg/ddn148.

Møller, Henrik Devitt. 2020. "Circle-Seq: Isolation and Sequencing of Chromosome-Derived Circular DNA Elements in Cells." In *DNA Electrophoresis: Methods and Protocols*, edited by Katsuhiro Hanada, 165–81. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-0323-9_15.

Møller, Henrik Devitt, Marghoob Mohiyuddin, Iñigo Prada-Luengo, M. Reza Sailani, Jens Frey Halling, Peter Plomgaard, Lasse Maretty, et al. 2018. "Circular DNA Elements of Chromosomal Origin Are Common in Healthy Human Somatic Tissue." *Nature Communications* 9 (1): 1069. https://doi.org/10.1038/s41467-018-03369-8.

Møller, Henrik D., Lance Parsons, Tue S. Jørgensen, David Botstein, and Birgitte Regenberg. 2015. "Extrachromosomal Circular DNA Is Common in Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 112 (24): E3114–22. https://doi.org/10.1073/pnas.1508825112.

Mootha, Vamsi K., Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. "PGC-1alpha-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes." *Nature Genetics* 34 (3): 267–73. https://doi.org/10.1038/ng1180.

Morison, Ian M., Joshua P. Ramsay, and Hamish G. Spencer. 2005. "A Census of Mammalian Imprinting." *Trends in Genetics: TIG* 21 (8): 457–65. https://doi.org/10.1016/j.tig.2005.06.008.

Morley, Michael, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, and Vivian G. Cheung. 2004. "Genetic Analysis of Genome-Wide Variation in Human Gene Expression." *Nature* 430 (7001): 743–47. https://doi.org/10.1038/nature02797.

Morton, Andrew R., Nergiz Dogan-Artun, Zachary J. Faber, Graham MacLeod, Cynthia F. Bartels, Megan S. Piazza, Kevin C. Allan, et al. 2019. "Functional Enhancers Shape Extrachromosomal Oncogene Amplifications." *Cell* 179 (6): 1330–41.e13. https://doi.org/10.1016/j.cell.2019.10.039.

Morton, Newton E. 2005. "Linkage Disequilibrium Maps and Association Mapping." *The Journal of Clinical Investigation* 115 (6): 1425–30. https://doi.org/10.1172/JCI25032.

Mossé, Yaël P., Marci Laudenslager, Luca Longo, Kristina A. Cole, Andrew Wood, Edward F. Attiyeh, Michael J. Laquaglia, et al. 2008. "Identification of ALK as a Major Familial Neuroblastoma Predisposition Gene." *Nature* 455 (7215): 930–35.

https://doi.org/10.1038/nature07261.
Murphy, E. V., Y. Zhang, W. Zhu, and J. Biggs. 1995. “The Human Glioma Pathogenesis-Related Protein Is Structurally Related to Plant Pathogenesis-Related Proteins and Its Gene Is Expressed Specifically in Brain Tumors.” *Gene* 159 (1): 131–35. https://doi.org/10.1016/0378-1119(95)00061-a.
Nature Methods. 2010. “E Pluribus Unum.” https://doi.org/10.1038/nmeth0510-331.
Nau, M. M., B. J. Brooks Jr, D. N. Carney, A. F. Gazdar, J. F. Battey, E. A. Sausville, and J. D. Minna. 1986. “Human Small-Cell Lung Cancers Show Amplification and Expression of the N-Myc Gene.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (4): 1092–96. https://doi.org/10.1073/pnas.83.4.1092.
Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, et al. 2011. “Tumour Evolution Inferred by Single-Cell Sequencing.” *Nature* 472 (7341): 90–94. https://doi.org/10.1038/nature09807.
Ng, Ray Kit, and J. B. Gurdon. 2008. “Epigenetic Memory of an Active Gene State Depends on Histone H3.3 Incorporation into Chromatin in the Absence of Transcription.” *Nature Cell Biology* 10 (1): 102–9. https://doi.org/10.1038/ncb1674.
Nguyen, Le B., Sharon J. Diskin, Mario Capasso, Kai Wang, Maura A. Diamond, Joseph Glessner, Cecilia Kim, et al. 2011. “Phenotype Restricted Genome-Wide Association Study Using a Gene-Centric Approach Identifies Three Low-Risk Neuroblastoma Susceptibility Loci.” *PLoS Genetics* 7 (3): e1002026. https://doi.org/10.1371/journal.pgen.1002026.
Nica, Alexandra C., Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. 2010. “Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations.” *PLoS Genetics* 6 (4): e1000895. https://doi.org/10.1371/journal.pgen.1000895.
Noguchi, T., K. Akiyama, M. Yokoyama, N. Kanda, T. Matsunaga, and Y. Nishi. 1996. “Amplification of a DEAD Box Gene (DDX1) with the MYCN Gene in Neuroblastomas as a Result of Cosegregation of Sequences Flanking the MYCN Locus.” *Genes, Chromosomes & Cancer* 15 (2): 129–33. https://doi.org/10.1002/(SICI)1098-2264(199602)15:2<129::AID-GCC8>3.0.CO;2-5.
Northcott, Paul A., Catherine Lee, Thomas Zichner, Adrian M. Stütz, Serap Erkek, Daisuke Kawauchi, David J. H. Shih, et al. 2014. “Enhancer Hijacking Activates GFI1 Family Oncogenes in Medulloblastoma.” *Nature* 511 (7510): 428–34. https://doi.org/10.1038/nature13379.
Nowell, P. C. 1976. “The Clonal Evolution of Tumor Cell Populations.” *Science* 194 (4260): 23–28. https://doi.org/10.1126/science.959840.
Nunberg, J. H., R. J. Kaufman, R. T. Schimke, G. Urlaub, and L. A. Chasin. 1978. “Amplified Dihydrofolate Reductase Genes Are Localized to a Homogeneously Staining Region of a Single Chromosome in a Methotrexate-Resistant Chinese Hamster Ovary Cell Line.” *Proceedings of the National Academy of Sciences of the United States of America* 75 (11): 5553–56. https://doi.org/10.1073/pnas.75.11.5553.
Oberthuer, André, Jessica Theissen, Frank Westermann, Barbara Hero, and Matthias Fischer. 2009. “Molecular Characterization and Classification of Neuroblastoma.” *Future Oncology* 5 (5): 625–39. https://doi.org/10.2217/fon.09.41.
Ohshima, Keiichi, Keiichi Hatakeyama, Takeshi Nagashima, Yuko Watanabe, Kaori Kanto, Yuki Doi, Tomomi Ide, et al. 2017. “Integrated Analysis of Gene Expression and Copy Number Identified Potential Cancer Driver Genes with Amplification-Dependent Overexpression in 1,454 Solid Tumors.” *Scientific Reports* 7 (1): 641. https://doi.org/10.1038/s41598-017-00219-3.
Okumura, K., R. Kiyama, and M. Oishi. 1987. “Sequence Analyses of Extrachromosomal

Sau3A and Related Family DNA: Analysis of Recombination in the Excision Event." *Nucleic Acids Research* 15 (18): 7477–89. https://doi.org/10.1093/nar/15.18.7477.
Oldridge, Derek A., Andrew C. Wood, Nina Weichert-Leahey, Ian Crimmins, Robyn Sussman, Cynthia Winter, Lee D. McDaniel, et al. 2015. "Genetic Predisposition to Neuroblastoma Mediated by a LMO1 Super-Enhancer Polymorphism." *Nature* 528 (7582): 418–21. https://doi.org/10.1038/nature15540.
Oldridge, E. E., H. F. Walker, M. J. Stower, M. S. Simms, V. M. Mann, A. T. Collins, D. Pellacani, and N. J. Maitland. 2013. "Retinoic Acid Represses Invasion and Stem Cell Phenotype by Induction of the Metastasis Suppressors RARRES1 and LXN." *Oncogenesis* 2 (April): e45. https://doi.org/10.1038/oncsis.2013.6.
Olsson, Maja, Stephan Beck, Per Kogner, Tommy Martinsson, and Helena Carén. 2016. "Genome-Wide Methylation Profiling Identifies Novel Methylated Genes in Neuroblastoma Tumors." *Epigenetics: Official Journal of the DNA Methylation Society* 11 (1): 74–84. https://doi.org/10.1080/15592294.2016.1138195.
Onengut-Gumuscu, Suna, Wei-Min Chen, Oliver Burren, Nick J. Cooper, Aaron R. Quinlan, Josyf C. Mychaleckyj, Emily Farber, et al. 2015. "Fine Mapping of Type 1 Diabetes Susceptibility Loci and Evidence for Colocalization of Causal Variants with Lymphoid Gene Enhancers." *Nature Genetics* 47 (4): 381–86. https://doi.org/10.1038/ng.3245.
Ong, Chin-Tong, and Victor G. Corces. 2011. "Enhancer Function: New Insights into the Regulation of Tissue-Specific Gene Expression." *Nature Reviews. Genetics* 12 (4): 283–93. https://doi.org/10.1038/nrg2957.
———. 2014. "CTCF: An Architectural Protein Bridging Genome Topology and Function." *Nature Reviews. Genetics* 15 (4): 234–46. https://doi.org/10.1038/nrg3663.
Ongen, Halit, Claus L. Andersen, Jesper B. Bramsen, Bodil Oster, Mads H. Rasmussen, Pedro G. Ferreira, Juan Sandoval, et al. 2014. "Putative Cis-Regulatory Drivers in Colorectal Cancer." *Nature* 512 (7512): 87–90. https://doi.org/10.1038/nature13602.
Ono, R., S. Kobayashi, H. Wagatsuma, K. Aisaka, T. Kohda, T. Kaneko-Ishino, and F. Ishino. 2001. "A Retrotransposon-Derived Gene, PEG10, Is a Novel Imprinted Gene Located on Human Chromosome 7q21." *Genomics* 73 (2): 232–37. https://doi.org/10.1006/geno.2001.6494.
Ormandy, Christopher J., Elizabeth A. Musgrove, Rina Hui, Roger J. Daly, and Robert L. Sutherland. 2003. "Cyclin D1, EMS1 and 11q13 Amplification in Breast Cancer." *Breast Cancer Research and Treatment* 78 (3): 323–35. https://doi.org/10.1023/a:1023033708204.
Ouwens, Klaasjan G., Rick Jansen, Michel G. Nivard, Jenny van Dongen, Maia J. Frieser, Jouke-Jan Hottenga, Wibowo Arindrarto, et al. 2020. "A Characterization of Cis- and Trans-Heritability of RNA-Seq-Based Gene Expression." *European Journal of Human Genetics: EJHG* 28 (2): 253–63. https://doi.org/10.1038/s41431-019-0511-5.
Pandey, Gaurav Kumar, Sanhita Mitra, Santhilal Subhash, Falk Hertwig, Meena Kanduri, Kankadeb Mishra, Susanne Fransson, et al. 2014. "The Risk-Associated Long Noncoding RNA NBAT-1 Controls Neuroblastoma Progression by Regulating Cell Proliferation and Neuronal Differentiation." *Cancer Cell* 26 (5): 722–37. https://doi.org/10.1016/j.ccell.2014.09.014.
Pant, Vinod, Sreenivasulu Kurukuti, Elena Pugacheva, Shaharum Shamsuddin, Piero Mariano, Rainer Renkawitz, Elena Klenova, Victor Lobanenkov, and Rolf Ohlsson. 2004. "Mutation of a Single CTCF Target Site within the H19 Imprinting Control Region Leads to Loss of Igf2 Imprinting and Complex Patterns of de Novo Methylation upon Maternal Inheritance." *Molecular and Cellular Biology* 24 (8): 3497–3504. https://doi.org/10.1128/mcb.24.8.3497-3504.2004.
Park, K. J., K. H. Shin, J. L. Ku, T. J. Cho, S. H. Lee, I. H. Choi, C. Phillipe, A. P. Monaco, D. E. Porter, and J. G. Park. 1999. "Germline Mutations in the EXT1 and EXT2 Genes in

Korean Patients with Hereditary Multiple Exostoses." *Journal of Human Genetics* 44 (4): 230–34. https://doi.org/10.1007/s100380050149.
Pastinen, Tomi. 2010. "Genome-Wide Allele-Specific Analysis: Insights into Regulatory Variation." *Nature Reviews. Genetics* 11 (8): 533–38. https://doi.org/10.1038/nrg2815.
Patwardhan, Rupali P., Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, et al. 2012. "Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo." *Nature Biotechnology* 30 (3): 265–70. https://doi.org/10.1038/nbt.2136.
Paulsen, Teressa, Pankaj Kumar, M. Murat Koseoglu, and Anindya Dutta. 2018. "Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells." *Trends in Genetics: TIG* 34 (4): 270–78. https://doi.org/10.1016/j.tig.2017.12.010.
Paulsen, Teressa, Pumoli Malapati, Rebeka Eki, Tarek Abbas, and Anindya Dutta. 2020. "EccDNA Formation Is Dependent on MMEJ, Repressed by c-NHEJ Pathway, and Stimulated by DNA Double-Strand Break." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.12.03.410480.
PCAWG Transcriptome Core Group, Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğlu, Nuno A. Fonseca, Yao He, André Kahles, et al. 2020. "Genomic Basis for RNA Alterations in Cancer." *Nature* 578 (7793): 129–36. https://doi.org/10.1038/s41586-020-1970-0.
Peifer, Martin, Falk Hertwig, Frederik Roels, Daniel Dreidax, Moritz Gartlgruber, Roopika Menon, Andrea Krämer, et al. 2015. "Telomerase Activation by Genomic Rearrangements in High-Risk Neuroblastoma." *Nature* 526 (7575): 700–704. https://doi.org/10.1038/nature14980.
Peters, Jo. 2014. "The Role of Genomic Imprinting in Biology and Disease: An Expanding View." *Nature Reviews. Genetics* 15 (8): 517–30. https://doi.org/10.1038/nrg3766.
Philippe, C., D. E. Porter, M. E. Emerton, D. E. Wells, A. H. Simpson, and A. P. Monaco. 1997. "Mutation Screening of the EXT1 and EXT2 Genes in Patients with Hereditary Multiple Exostoses." *American Journal of Human Genetics* 61 (3): 520–28. https://doi.org/10.1086/515505.
Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, et al. 1998. "High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays." *Nature Genetics* 20 (2): 207–11. https://doi.org/10.1038/2524.
Pisitkun, Prapaporn, Jonathan A. Deane, Michael J. Difilippantonio, Tatyana Tarasenko, Anne B. Satterthwaite, and Silvia Bolland. 2006. "Autoreactive B Cell Responses to RNA-Related Antigens due to TLR7 Gene Duplication." *Science* 312 (5780): 1669–72. https://doi.org/10.1126/science.1124978.
Plagnol, Vincent, Elif Uz, Chris Wallace, Helen Stevens, David Clayton, Tayfun Ozcelik, and John A. Todd. 2008. "Extreme Clonality in Lymphoblastoid Cell Lines with Implications for Allele Specific Expression Analyses." *PloS One* 3 (8): e2966. https://doi.org/10.1371/journal.pone.0002966.
Pollard, Katherine S., David Serre, Xu Wang, Heng Tao, Elin Grundberg, Thomas J. Hudson, Andrew G. Clark, and Kelly Frazer. 2008. "A Genome-Wide Approach to Identifying Novel-Imprinted Genes." *Human Genetics* 122 (6): 625–34. https://doi.org/10.1007/s00439-007-0440-1.
Pombo, Ana, and Niall Dillon. 2015. "Three-Dimensional Genome Architecture: Players and Mechanisms." *Nature Reviews. Molecular Cell Biology* 16 (4): 245–57. https://doi.org/10.1038/nrm3965.
Pontual, Loïc de, Delphine Trochet, Franck Bourdeaut, Sophie Thomas, Heather Etchevers, Agnes Chompret, Véronique Minard, et al. 2007. "Methylation-Associated PHOX2B Gene Silencing Is a Rare Event in Human Neuroblastoma." *European Journal of Cancer*

43 (16): 2366–72. https://doi.org/10.1016/j.ejca.2007.07.016.
Popova, Tatiana, Elodie Manié, Dominique Stoppa-Lyonnet, Guillem Rigaill, Emmanuel Barillot, and Marc Henri Stern. 2009. “Genome Alteration Print (GAP): A Tool to Visualize and Mine Complex Cancer Genomic Profiles Obtained by SNP Arrays.” *Genome Biology* 10 (11): R128. https://doi.org/10.1186/gb-2009-10-11-r128.
Poremba, C., H. Willenbring, B. Hero, H. Christiansen, K. L. Schäfer, C. Brinkschmidt, H. Jürgens, W. Böcker, and B. Dockhorn-Dworniczak. 1999. “Telomerase Activity Distinguishes between Neuroblastomas with Good and Poor Prognosis.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 10 (6): 715–21. https://doi.org/10.1023/a:1008333500733.
Potapova, Tamara A., Jin Zhu, and Rong Li. 2013. “Aneuploidy and Chromosomal Instability: A Vicious Cycle Driving Cellular Evolution and Cancer Genome Chaos.” *Cancer Metastasis Reviews* 32 (3-4): 377–89. https://doi.org/10.1007/s10555-013-9436-6.
Prawitt, Dirk, Thorsten Enklaar, Barbara Gärtner-Rupprecht, Christian Spangenberg, Monika Oswald, Ekkehart Lausch, Peter Schmidtke, et al. 2005. “Microdeletion of Target Sites for Insulator Protein CTCF in a Chromosome 11p15 Imprinting Center in Beckwith-Wiedemann Syndrome and Wilms’ Tumor.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (11): 4085–90. https://doi.org/10.1073/pnas.0500037102.
Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.
Pritchard, J., and J. A. Hickman. 1994. “Why Does Stage 4s Neuroblastoma Regress Spontaneously?” *The Lancet* 344 (8926): 869–70. https://doi.org/10.1016/s0140-6736(94)92834-7.
Pritchard, J. K., and N. A. Rosenberg. 1999. “Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies.” *American Journal of Human Genetics* 65 (1): 220–28. https://doi.org/10.1086/302449.
Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2005. “NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins.” *Nucleic Acids Research* 33 (Database issue): D501–4. https://doi.org/10.1093/nar/gki025.
Przytycki, Pawel F., and Mona Singh. 2020. “Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations.” *Cell Systems* 10 (2): 193–203.e4. https://doi.org/10.1016/j.cels.2020.01.002.
Puente, Xose S., Silvia Beà, Rafael Valdés-Mas, Neus Villamor, Jesús Gutiérrez-Abril, José I. Martín-Subero, Marta Munar, et al. 2015. “Non-Coding Recurrent Mutations in Chronic Lymphocytic Leukaemia.” *Nature* 526 (7574): 519–24. https://doi.org/10.1038/nature14666.
Pugh, Trevor J., Olena Morozova, Edward F. Attiyeh, Shahab Asgharzadeh, Jun S. Wei, Daniel Auclair, Scott L. Carter, et al. 2013. “The Genetic Landscape of High-Risk Neuroblastoma.” *Nature Genetics* 45 (3): 279–84. https://doi.org/10.1038/ng.2529.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.
Quang, Daniel, and Xiaohui Xie. 2019. “FactorNet: A Deep Learning Framework for Predicting Cell Type Specific Transcription Factor Binding from Nucleotide-Resolution Sequential Data.” *Methods* 166 (August): 40–47. https://doi.org/10.1016/j.ymeth.2019.03.020.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Quinn, Michael C. J., Abdelali Filali-Mouhim, Diane M. Provencher, Anne-Marie Mes-Masson, and Patricia N. Tonin. 2009. “Reprogramming of the Transcriptome in a Novel Chromosome 3 Transfer Tumor Suppressor Ovarian Cancer Cell Line Model Affected Molecular Networks That Are Characteristic of Ovarian Cancer.” *Molecular Carcinogenesis* 48 (7): 648–61. https://doi.org/10.1002/mc.20511.

Raabe, E. H., M. Laudenslager, C. Winter, N. Wasserman, K. Cole, M. LaQuaglia, D. J. Maris, Y. P. Mosse, and J. M. Maris. 2008. “Prevalence and Functional Consequence of PHOX2B Mutations in Neuroblastoma.” *Oncogene* 27 (4): 469–76. https://doi.org/10.1038/sj.onc.1210659.

Radloff, R., W. Bauer, and J. Vinograd. 1967. “A Dye-Buoyant-Density Method for the Detection and Isolation of Closed Circular Duplex DNA: The Closed Circular DNA in HeLa Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 57 (5): 1514–21. https://doi.org/10.1073/pnas.57.5.1514.

Raine, Keiran M., Peter Van Loo, David C. Wedge, David Jones, Andrew Menzies, Adam P. Butler, Jon W. Teague, Patrick Tarpey, Serena Nik-Zainal, and Peter J. Campbell. 2016. “ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 56 (December): 15.9.1–15.9.17. https://doi.org/10.1002/cpbi.17.

Ram Kumar, Ram Mohan, and Nina Felice Schor. 2018. “Methylation of DNA and Chromatin as a Mechanism of Oncogenesis and Therapeutic Target in Neuroblastoma.” *Oncotarget* 9 (31): 22184–93. https://doi.org/10.18632/oncotarget.25084.

Rasmussen, Markus, Magnus Sundström, Hanna Göransson Kultima, Johan Botling, Patrick Micke, Helgi Birgisson, Bengt Glimelius, and Anders Isaksson. 2011. “Allele-Specific Copy Number Analysis of Tumor Samples with Aneuploidy and Tumor Heterogeneity.” *Genome Biology* 12 (10): R108. https://doi.org/10.1186/gb-2011-12-10-r108.

Rausch, Tobias, David T. W. Jones, Marc Zapatka, Adrian M. Stütz, Thomas Zichner, Joachim Weischenfeldt, Natalie Jäger, et al. 2012. “Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations.” *Cell* 148 (1-2): 59–71. https://doi.org/10.1016/j.cell.2011.12.013.

Reinius, Björn, and Rickard Sandberg. 2015. “Random Monoallelic Expression of Autosomal Genes: Stochastic Transcription and Allele-Level Regulation.” *Nature Reviews. Genetics* 16 (11): 653–64. https://doi.org/10.1038/nrg3888.

Ren, Chengzhen, Likun Li, Guang Yang, Terry L. Timme, Alexei Goltsov, Chenghui Ren, Xiaorong Ji, et al. 2004. “RTVP-1, a Tumor Suppressor Inactivated by Methylation in Prostate Cancer.” *Cancer Research* 64 (3): 969–76. https://doi.org/10.1158/0008-5472.can-03-2592.

Ren, Chengzhen, Cheng-Hui Ren, Likun Li, Alexei A. Goltsov, and Timothy C. Thompson. 2006. “Identification and Characterization of RTVP1/GLIPR1-like Genes, a Novel p53 Target Gene Cluster.” *Genomics* 88 (2): 163–72. https://doi.org/10.1016/j.ygeno.2006.03.021.

Ren, Gang, Wenfei Jin, Kairong Cui, Joseph Rodrigez, Gangqing Hu, Zhiying Zhang, Daniel R. Larson, and Keji Zhao. 2017. “CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression.” *Molecular Cell* 67 (6): 1049–58.e6. https://doi.org/10.1016/j.molcel.2017.08.026.

Rheinbay, Esther, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, et al. 2020. “Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes.” *Nature* 578 (7793): 102–11.

https://doi.org/10.1038/s41586-020-1965-x.
Rich, T., P. Chen, F. Furman, N. Huynh, and M. A. Israel. 1996. “RTVP-1, a Novel Human Gene with Sequence Similarity to Genes of Diverse Species, Is Expressed in Tumor Cell Lines of Glial but Not Neuronal Origin.” *Gene* 180 (1-2): 125–30. https://doi.org/10.1016/s0378-1119(96)00431-3.
Riordan, Jesse D., and Adam J. Dupuy. 2013. “Domesticated Transposable Element Gene Products in Human Cancer.” *Mobile Genetic Elements* 3 (5): e26693. https://doi.org/10.4161/mge.26693.
Riordan, Jesse D., Vincent W. Keng, Barbara R. Tschida, Todd E. Scheetz, Jason B. Bell, Kelly M. Podetz-Pedersen, Catherine D. Moser, et al. 2013. “Identification of rtl1, a Retrotransposon-Derived Imprinted Gene, as a Novel Driver of Hepatocarcinogenesis.” *PLoS Genetics* 9 (4): e1003441. https://doi.org/10.1371/journal.pgen.1003441.
Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. “Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing.” *Nature Methods* 4 (8): 651–57. https://doi.org/10.1038/nmeth1068.
Robyr, Daniel, Siavash K. Kurdistani, and Michael Grunstein. 2003. “Analysis of Genome-Wide Histone Acetylation State and Enzyme Binding Using DNA Microarrays.” In *Methods in Enzymology*, 376:289–304. Academic Press. https://doi.org/10.1016/S0076-6879(03)76019-4.
Rocha, Simao Teixeira da, Carol A. Edwards, Mitsuteru Ito, Tsutomu Ogata, and Anne C. Ferguson-Smith. 2008. “Genomic Imprinting at the Mammalian Dlk1-Dio3 Domain.” *Trends in Genetics: TIG* 24 (6): 306–16. https://doi.org/10.1016/j.tig.2008.03.011.
Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, et al. 2011. “AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework.” *Molecular Systems Biology* 7 (August): 522. https://doi.org/10.1038/msb.2011.54.
Rubin, Adam J., Brook C. Barajas, Mayra Furlan-Magaril, Vanessa Lopez-Pajares, Maxwell R. Mumbach, Imani Howard, Daniel S. Kim, et al. 2017. “Lineage-Specific Dynamic and Pre-Established Enhancer--Promoter Contacts Cooperate in Terminal Differentiation.” *Nature Genetics* 49 (10): 1522. https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/ng.3935.pdf%3Forigin%3Dppub&casa_token=3vGdyJHaUwEAAAAA:MyqlF22MJZVa8N_8PMRKi3Dfw2o8ZWUfDZnicX_OD8Qet4M_5zKSM9ffeJs8G60EwUMgxR8MY4cJ4H3E.
Ruiz, J. C., and G. M. Wahl. 1990. “Chromosomal Destabilization during Gene Amplification.” *Molecular and Cellular Biology* 10 (6): 3056–66. https://doi.org/10.1128/mcb.10.6.3056.
Russell, Mike R., Annalise Penikis, Derek A. Oldridge, Juan R. Alvarez-Dominguez, Lee McDaniel, Maura Diamond, Olivia Padovan, et al. 2015. “CASC15-S Is a Tumor Suppressor lncRNA at the 6p22 Neuroblastoma Susceptibility Locus.” *Cancer Research* 75 (15): 3155–66. https://doi.org/10.1158/0008-5472.CAN-14-3613.
Sack, Laura Magill, Teresa Davoli, Mamie Z. Li, Yuyang Li, Qikai Xu, Kamila Naxerova, Eric C. Wooten, et al. 2018. “Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns.” *Cell* 173 (2): 499–514.e23. https://doi.org/10.1016/j.cell.2018.02.037.
Saint-André, Violaine, Alexander J. Federation, Charles Y. Lin, Brian J. Abraham, Jessica Reddy, Tong Ihn Lee, James E. Bradner, and Richard A. Young. 2016. “Models of Human Core Transcriptional Regulatory Circuitries.” *Genome Research* 26 (3): 385–96. https://doi.org/10.1101/gr.197590.115.
Sakatani, T., M. Wei, M. Katoh, C. Okita, D. Wada, K. Mitsuya, M. Meguro, et al. 2001. “Epigenetic Heterogeneity at Imprinted Loci in Normal Populations.” *Biochemical and*

Biophysical Research Communications 283 (5): 1124–30. https://doi.org/10.1006/bbrc.2001.4916.
Sanchis-Juan, Alba, Jonathan Stephens, Courtney E. French, Nicholas Gleadall, Karyn Mégy, Christopher Penkett, Olga Shamardina, et al. 2018. “Complex Structural Variants in Mendelian Disorders: Identification and Breakpoint Resolution Using Short- and Long-Read Genome Sequencing.” *Genome Medicine* 10 (1): 95. https://doi.org/10.1186/s13073-018-0606-6.
Sansregret, Laurent, and Charles Swanton. 2017. “The Role of Aneuploidy in Cancer Evolution.” *Cold Spring Harbor Perspectives in Medicine* 7 (1). https://doi.org/10.1101/cshperspect.a028373.
Sasaki, H., P. A. Jones, J. R. Chaillet, A. C. Ferguson-Smith, S. C. Barton, W. Reik, and M. A. Surani. 1992. “Parental Imprinting: Potentially Active Chromatin of the Repressed Maternal Allele of the Mouse Insulin-like Growth Factor II (Igf2) Gene.” *Genes & Development* 6 (10): 1843–56. https://doi.org/10.1101/gad.6.10.1843.
Sathirapongsasuti, Jarupon Fah, Hane Lee, Basil A. J. Horst, Georg Brunner, Alistair J. Cochran, Scott Binder, John Quackenbush, and Stanley F. Nelson. 2011. “Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV.” *Bioinformatics* 27 (19): 2648–54. https://doi.org/10.1093/bioinformatics/btr462.
Saxonov, Serge, Paul Berg, and Douglas L. Brutlag. 2006. “A Genome-Wide Analysis of CpG Dinucleotides in the Human Genome Distinguishes Two Distinct Classes of Promoters.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (5): 1412–17. https://doi.org/10.1073/pnas.0510310103.
Schadt, Eric E., Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y. Lum, Andrew Kasarskis, et al. 2008. “Mapping the Genetic Architecture of Gene Expression in Human Liver.” *PLoS Biology* 6 (5): e107. https://doi.org/10.1371/journal.pbio.0060107.
Schadt, Eric E., Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, et al. 2003. “Genetics of Gene Expression Surveyed in Maize, Mouse and Man.” *Nature* 422 (6929): 297–302. https://doi.org/10.1038/nature01434.
Schimke, R. T., R. J. Kaufman, F. W. Alt, and R. F. Kellems. 1978. “Gene Amplification and Drug Resistance in Cultured Murine Cells.” *Science* 202 (4372): 1051–55. https://doi.org/10.1126/science.715457.
Schmiedel, Benjamin Joachim, Grégory Seumois, Daniela Samaniego-Castruita, Justin Cayford, Veronique Schulten, Lukas Chavez, Ferhat Ay, Alessandro Sette, Bjoern Peters, and Pandurangan Vijayanand. 2016. “17q21 Asthma-Risk Variants Switch CTCF Binding and Regulate IL-2 Production by T Cells.” *Nature Communications* 7 (November): 13426. https://doi.org/10.1038/ncomms13426.
Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer-Promoter Contacts in Gene Expression Control.” *Nature Reviews. Genetics* 20 (8): 437–55. https://doi.org/10.1038/s41576-019-0128-0.
Schones, Dustin E., Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. 2008. “Dynamic Regulation of Nucleosome Positioning in the Human Genome.” *Cell* 132 (5): 887–98. https://doi.org/10.1016/j.cell.2008.02.022.
Schramm, Alexander, Johannes Köster, Yassen Assenov, Kristina Althoff, Martin Peifer, Ellen Mahlow, Andrea Odersky, et al. 2015. “Mutational Dynamics between Primary and Relapse Neuroblastomas.” *Nature Genetics* 47 (8): 872–77. https://doi.org/10.1038/ng.3349.
Schröck, E., S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, et al. 1996. “Multicolor Spectral Karyotyping of Human Chromosomes.” *Science* 273 (5274): 494–97. https://doi.org/10.1126/science.273.5274.494.

Schüle, Birgitt, Karen N. McFarland, Kelsey Lee, Yu-Chih Tsai, Khanh-Dung Nguyen, Chao Sun, Mei Liu, et al. 2017. “Parkinson’s Disease Associated with Pure ATXN10 Repeat Expansion.” *NPJ Parkinson’s Disease* 3 (September): 27. https://doi.org/10.1038/s41531-017-0029-x.

Schulte, Johannes H., Hagen S. Bachmann, Bent Brockmeyer, Katleen Depreter, André Oberthür, Sandra Ackermann, Yvonne Kahlert, et al. 2011. “High ALK Receptor Tyrosine Kinase Expression Supersedes ALK Mutation as a Determining Factor of an Unfavorable Phenotype in Primary Neuroblastoma.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 17 (15): 5082–92. https://doi.org/10.1158/1078-0432.CCR-10-2809.

Schuster-Böckler, Benjamin, Donald Conrad, and Alex Bateman. 2010. “Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions.” *PloS One* 5 (3): e9474. https://doi.org/10.1371/journal.pone.0009474.

Schwab, M., K. Alitalo, K. H. Klempnauer, H. E. Varmus, J. M. Bishop, F. Gilbert, G. Brodeur, M. Goldstein, and J. Trent. 1983. “Amplified DNA with Limited Homology to Myc Cellular Oncogene Is Shared by Human Neuroblastoma Cell Lines and a Neuroblastoma Tumour.” *Nature* 305 (5931): 245–48. https://doi.org/10.1038/305245a0.

Schwab, M., and L. C. Amler. 1990. “Amplification of Cellular Oncogenes: A Predictor of Clinical Outcome in Human Cancer.” *Genes, Chromosomes & Cancer* 1 (3): 181–93. https://doi.org/10.1002/gcc.2870010302.

Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. “Global Quantification of Mammalian Gene Expression Control.” *Nature* 473 (7347): 337–42. https://doi.org/10.1038/nature10098.

Schwartz, Brian E., and Kami Ahmad. 2005. “Transcriptional Activation Triggers Deposition and Removal of the Histone Variant H3.3.” *Genes & Development* 19 (7): 804–14. https://doi.org/10.1101/gad.1259805.

Schwartzentruber, Jeremy, Andrey Korshunov, Xiao-Yang Liu, David T. W. Jones, Elke Pfaff, Karine Jacob, Dominik Sturm, et al. 2012. “Driver Mutations in Histone H3.3 and Chromatin Remodelling Genes in Paediatric Glioblastoma.” *Nature* 482 (7384): 226–31. https://doi.org/10.1038/nature10833.

Sebat, Jonathan, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, et al. 2004. “Large-Scale Copy Number Polymorphism in the Human Genome.” *Science* 305 (5683): 525–28. https://doi.org/10.1126/science.1098918.

Seeger, R. C., G. M. Brodeur, H. Sather, A. Dalton, S. E. Siegel, K. Y. Wong, and D. Hammond. 1985. “Association of Multiple Copies of the N-Myc Oncogene with Rapid Progression of Neuroblastomas.” *The New England Journal of Medicine* 313 (18): 1111–16. https://doi.org/10.1056/NEJM198510313131802.

Sekita, Yoichi, Hirotaka Wagatsuma, Kenji Nakamura, Ryuichi Ono, Masayo Kagami, Noriko Wakisaka, Toshiaki Hino, et al. 2008. “Role of Retrotransposon-Derived Imprinted Gene, Rtl1, in the Feto-Maternal Interface of Mouse Placenta.” *Nature Genetics* 40 (2): 243–48. https://doi.org/10.1038/ng.2007.51.

Selmecki, Anna M., Yosef E. Maruvka, Phillip A. Richmond, Marie Guillet, Noam Shoresh, Amber L. Sorenson, Subhajyoti De, et al. 2015. “Polyploidy Can Drive Rapid Adaptation in Yeast.” *Nature* 519 (7543): 349–52. https://doi.org/10.1038/nature14187.

Shabalin, Andrey A. 2012. “Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations.” *Bioinformatics* 28 (10): 1353–58. https://doi.org/10.1093/bioinformatics/bts163.

Shaikh, Tamim H. 2017. “Copy Number Variation Disorders.” *Current Genetic Medicine Reports* 5 (4): 183–90. https://doi.org/10.1007/s40142-017-0129-2.

Shao, Xin, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan.

2019. “Copy Number Variation Is Highly Correlated with Differential Gene Expression: A Pan-Cancer Study.” *BMC Medical Genetics* 20 (1): 175. https://doi.org/10.1186/s12881-019-0909-5.

Sharon, Eilon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. 2012. “Inferring Gene Regulatory Logic from High-Throughput Measurements of Thousands of Systematically Designed Promoters.” *Nature Biotechnology* 30 (6): 521–30. https://doi.org/10.1038/nbt.2205.

Shen, Ronglai, and Venkatraman E. Seshan. 2016. “FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool for High-Throughput DNA Sequencing.” *Nucleic Acids Research* 44 (16): e131–e131. https://doi.org/10.1093/nar/gkw520.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. “dbSNP: The NCBI Database of Genetic Variation.” *Nucleic Acids Research* 29 (1): 308–11. https://doi.org/10.1093/nar/29.1.308.

Shibata, Yoshiyuki, Pankaj Kumar, Ryan Layer, Smaranda Willcox, Jeffrey R. Gagan, Jack D. Griffith, and Anindya Dutta. 2012. “Extrachromosomal microDNAs and Chromosomal Microdeletions in Normal Tissues.” *Science* 336 (6077): 82–86. https://doi.org/10.1126/science.1213307.

Shi, Junwei, Warren A. Whyte, Cinthya J. Zepeda-Mendoza, Joseph P. Milazzo, Chen Shen, Jae-Seok Roe, Jessica L. Minder, et al. 2013. “Role of SWI/SNF in Acute Leukemia Maintenance and Enhancer-Mediated Myc Regulation.” *Genes & Development* 27 (24): 2648–62. https://doi.org/10.1101/gad.232710.113.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. “Transcriptional Enhancers: From Properties to Genome-Wide Predictions.” *Nature Reviews. Genetics* 15 (4): 272–86. https://doi.org/10.1038/nrg3682.

Shoshani, Ofer, Simon F. Brunner, Rona Yaeger, Peter Ly, Yael Nechemia-Arbely, Dong Hyun Kim, Rongxin Fang, et al. 2020. “Chromothripsis Drives the Evolution of Gene Amplification in Cancer.” *Nature*, December. https://doi.org/10.1038/s41586-020-03064-z.

Siegl-Cachedenier, Irene, Purificación Muñoz, Juana M. Flores, Peter Klatt, and María A. Blasco. 2007. “Deficient Mismatch Repair Improves Organismal Fitness and Survival of Mice with Dysfunctional Telomeres.” *Genes & Development* 21 (17): 2234–47. https://doi.org/10.1101/gad.430107.

Siersbæk, Rasmus, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen, Emilie Kristine Bagge, Jonathan Cairns, Steven William Wingett, et al. 2017. “Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation.” *Molecular Cell* 66 (3): 420–35.e5. https://doi.org/10.1016/j.molcel.2017.04.010.

Sieverling, Lina, Chen Hong, Sandra D. Koser, Philip Ginsbach, Kortine Kleinheinz, Barbara Hutter, Delia M. Braun, et al. 2020. “Genomic Footprints of Activated Telomere Maintenance Mechanisms in Cancer.” *Nature Communications* 11 (1): 733. https://doi.org/10.1038/s41467-019-13824-9.

Simon, Jeffrey A., and Robert E. Kingston. 2009. “Mechanisms of Polycomb Gene Silencing: Knowns and Unknowns.” *Nature Reviews. Molecular Cell Biology* 10 (10): 697–708. https://doi.org/10.1038/nrm2763.

Sinclair, D. A., and L. Guarente. 1997. “Extrachromosomal rDNA Circles--a Cause of Aging in Yeast.” *Cell* 91 (7): 1033–42. https://doi.org/10.1016/s0092-8674(00)80493-6.

Smemo, Scott, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. “Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3.” *Nature* 507 (7492): 371–75. https://doi.org/10.1038/nature13138.

Smith, Malcolm A., Nita L. Seibel, Sean F. Altekruse, Lynn A. G. Ries, Danielle L. Melbert,

Maura O'Leary, Franklin O. Smith, and Gregory H. Reaman. 2010. "Outcomes for Children and Adolescents with Cancer: Challenges for the Twenty-First Century." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 28 (15): 2625–34. https://doi.org/10.1200/JCO.2009.27.0421.

Solimini, Nicole L., Qikai Xu, Craig H. Mermel, Anthony C. Liang, Michael R. Schlabach, Ji Luo, Anna E. Burrows, et al. 2012. "Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential." *Science* 337 (6090): 104–9. https://doi.org/10.1126/science.1219580.

Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, and P. Lichter. 1997. "Matrix-Based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances." *Genes, Chromosomes & Cancer* 20 (4): 399–407. https://www.ncbi.nlm.nih.gov/pubmed/9408757.

Song, Lingyun, and Gregory E. Crawford. 2010. "DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells." *Cold Spring Harbor Protocols* 2010 (2): db.prot5384. https://doi.org/10.1101/pdb.prot5384.

Sparago, Angela, Flavia Cerrato, Maria Vernucci, Giovanni Battista Ferrero, Margherita Cirillo Silengo, and Andrea Riccio. 2004. "Microdeletions in the Human H19 DMR Result in Loss of IGF2 Imprinting and Beckwith-Wiedemann Syndrome." *Nature Genetics* 36 (9): 958–60. https://doi.org/10.1038/ng1410.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. "Structural Variation in the 3D Genome." *Nature Reviews. Genetics* 19 (7): 453–67. https://doi.org/10.1038/s41576-018-0007-0.

Spitz, Ruediger, Barbara Hero, Thorsten Simon, and Frank Berthold. 2006. "Loss in Chromosome 11q Identifies Tumors with Increased Risk for Metastatic Relapses in Localized and 4S Neuroblastoma." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 12 (11 Pt 1): 3368–73. https://doi.org/10.1158/1078-0432.CCR-05-2495.

Squire, J. A., P. S. Thorner, S. Weitzman, J. D. Maggi, P. Dirks, J. Doyle, M. Hale, and R. Godbout. 1995. "Co-Amplification of MYCN and a DEAD Box Gene (DDX1) in Primary Neuroblastoma." *Oncogene* 10 (7): 1417–22. https://www.ncbi.nlm.nih.gov/pubmed/7731693.

Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. 2012. "Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses." *Nature Protocols* 7 (3): 500. https://www.nature.com/articles/nprot.2011.457.pdf?origin=ppub.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. "A New Statistical Method for Haplotype Reconstruction from Population Data." *American Journal of Human Genetics* 68 (4): 978–89. https://doi.org/10.1086/319501.

Stern, C. 1943. "THE HARDY-WEINBERG LAW." *Science* 97 (2510): 137–38. https://doi.org/10.1126/science.97.2510.137.

Stiller, C. A., and D. M. Parkin. 1992. "International Variations in the Incidence of Neuroblastoma." *International Journal of Cancer. Journal International Du Cancer* 52 (4): 538–43. https://doi.org/10.1002/ijc.2910520407.

Storlazzi, Clelia Tiziana, Thoas Fioretos, Cecilia Surace, Angelo Lonoce, Angela Mastrorilli, Bodil Strömbeck, Pietro D'Addabbo, et al. 2006. "MYC-Containing Double Minutes in Hematologic Malignancies: Evidence in Favor of the Episome Model and Exclusion of MYC as the Target Gene." *Human Molecular Genetics* 15 (6): 933–42. https://doi.org/10.1093/hmg/ddl010.

Stranger, Barbara E., Matthew S. Forrest, Andrew G. Clark, Mark J. Minichiello, Samuel Deutsch, Robert Lyle, Sarah Hunt, et al. 2005. "Genome-Wide Associations of Gene

Expression Variation in Humans." *PLoS Genetics* 1 (6): e78. https://doi.org/10.1371/journal.pgen.0010078.
Stranger, Barbara E., Matthew S. Forrest, Mark Dunning, Catherine E. Ingle, Claude Beazley, Natalie Thorne, Richard Redon, et al. 2007. "Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes." *Science* 315 (5813): 848–53. https://doi.org/10.1126/science.1136678.
Stranger, Barbara E., Stephen B. Montgomery, Antigone S. Dimas, Leopold Parts, Oliver Stegle, Catherine E. Ingle, Magda Sekowska, et al. 2012. "Patterns of Cis Regulatory Variation in Diverse Human Populations." *PLoS Genetics* 8 (4): e1002639. https://doi.org/10.1371/journal.pgen.1002639.
Stutterheim, Janine, Annemieke Gerritsen, Lily Zappeij-Kannegieter, Ilona Kleijn, Rob Dee, Lotty Hooft, Max M. van Noesel, et al. 2008. "PHOX2B Is a Novel and Specific Marker for Minimal Residual Disease Testing in Neuroblastoma." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 26 (33): 5443–49. https://doi.org/10.1200/JCO.2007.13.6531.
Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.
Sun, F. L., W. L. Dean, G. Kelsey, N. D. Allen, and W. Reik. 1997. "Transactivation of Igf2 in a Mouse Model of Beckwith-Wiedemann Syndrome." *Nature* 389 (6653): 809–15. https://doi.org/10.1038/39797.
Talevich, Eric, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. 2016. "CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing." *PLoS Computational Biology* 12 (4): e1004873. https://doi.org/10.1371/journal.pcbi.1004873.
Tanner, M. M., M. Tirkkonen, A. Kallioniemi, J. Isola, T. Kuukasjärvi, C. Collins, D. Kowbel, et al. 1996. "Independent Amplification and Frequent Co-Amplification of Three Nonsyntenic Regions on the Long Arm of Chromosome 20 in Human Breast Cancer." *Cancer Research* 56 (15): 3441–45. https://www.ncbi.nlm.nih.gov/pubmed/8758909.
Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. "COSMIC: The Catalogue Of Somatic Mutations In Cancer." *Nucleic Acids Research* 47 (D1): D941–47. https://doi.org/10.1093/nar/gky1015.
Taub, R., I. Kirsch, C. Morton, G. Lenoir, D. Swan, S. Tronick, S. Aaronson, and P. Leder. 1982. "Translocation of the c-Myc Gene into the Immunoglobulin Heavy Chain Locus in Human Burkitt Lymphoma and Murine Plasmacytoma Cells." *Proceedings of the National Academy of Sciences of the United States of America* 79 (24): 7837–41. https://doi.org/10.1073/pnas.79.24.7837.
Teitz, T., T. Wei, M. B. Valentine, E. F. Vanin, J. Grenet, V. A. Valentine, F. G. Behm, A. T. Look, J. M. Lahti, and V. J. Kidd. 2000. "Caspase 8 Is Deleted or Silenced Preferentially in Childhood Neuroblastomas with Amplification of MYCN." *Nature Medicine* 6 (5): 529–35. https://doi.org/10.1038/75007.
Thomas, Mary C., and Cheng-Ming Chiang. 2006. "The General Transcription Machinery and General Cofactors." *Critical Reviews in Biochemistry and Molecular Biology* 41 (3): 105–78. https://doi.org/10.1080/10409230600648736.
Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. "The Accessible Chromatin Landscape of the Human Genome." *Nature* 489 (7414): 75–82. https://doi.org/10.1038/nature11232.
Tomlins, Scott A., Daniel R. Rhodes, Sven Perner, Saravana M. Dhanasekaran, Rohit

Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, et al. 2005. “Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer.” *Science* 310 (5748): 644–48. https://doi.org/10.1126/science.1117679.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics* 25 (9): 1105–11. https://doi.org/10.1093/bioinformatics/btp120.

Trochet, Delphine, Franck Bourdeaut, Isabelle Janoueix-Lerosey, Anne Deville, Loïc de Pontual, Gudrun Schleiermacher, Carole Coze, et al. 2004. “Germline Mutations of the Paired–Like Homeobox 2B (PHOX2B) Gene in Neuroblastoma.” *The American Journal of Human Genetics*. https://doi.org/10.1086/383253.

Trojer, Patrick, Guohong Li, Robert J. Sims 3rd, Alejandro Vaquero, Nagesh Kalakonda, Piernicola Boccuni, Donghoon Lee, et al. 2007. “L3MBTL1, a Histone-Methylation-Dependent Chromatin Lock.” *Cell* 129 (5): 915–28. https://doi.org/10.1016/j.cell.2007.03.048.

Trynka, Gosia, Cynthia Sandor, Buhm Han, Han Xu, Barbara E. Stranger, X. Shirley Liu, and Soumya Raychaudhuri. 2013. “Chromatin Marks Identify Critical Cell Types for Fine Mapping Complex Trait Variants.” *Nature Genetics* 45 (2): 124–30. https://doi.org/10.1038/ng.2504.

Tuch, Brian B., Rebecca R. Laborde, Xing Xu, Jian Gu, Christina B. Chung, Cinna K. Monighetti, Sarah J. Stanley, et al. 2010. “Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0009317.

Turner, Kristen M., Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, et al. 2017. “Extrachromosomal Oncogene Amplification Drives Tumour Evolution and Genetic Heterogeneity.” *Nature* 543 (7643): 122–25. https://doi.org/10.1038/nature21356.

Turro, Ernest, Shu-Yi Su, Ângela Gonçalves, Lachlan J. M. Coin, Sylvia Richardson, and Alex Lewin. 2011. “Haplotype and Isoform Specific Expression Estimation Using Multi-Mapping RNA-Seq Reads.” *Genome Biology* 12 (2): R13. https://doi.org/10.1186/gb-2011-12-2-r13.

Urahama, Takashi, Akihito Harada, Kazumitsu Maehara, Naoki Horikoshi, Koichi Sato, Yuko Sato, Koji Shiraishi, et al. 2016. “Histone H3.5 Forms an Unstable Nucleosome and Accumulates around Transcription Start Sites in Human Testis.” *Epigenetics & Chromatin* 9 (January): 2. https://doi.org/10.1186/s13072-016-0051-y.

Valentijn, Linda J., Jan Koster, Danny A. Zwijnenburg, Nancy E. Hasselt, Peter van Sluis, Richard Volckmann, Max M. van Noesel, et al. 2015. “TERT Rearrangements Are Frequent in Neuroblastoma and Identify Aggressive Tumors.” *Nature Genetics* 47 (12): 1411–14. https://doi.org/10.1038/ng.3438.

VanDevanter, D. R., V. D. Piaskowski, J. T. Casper, E. C. Douglass, and D. D. Von Hoff. 1990. “Ability of Circular Extrachromosomal DNA Molecules to Carry Amplified MYCN Proto-Oncogenes in Human Neuroblastomas in Vivo.” *Journal of the National Cancer Institute* 82 (23): 1815–21. https://doi.org/10.1093/jnci/82.23.1815.

Van Limpt, Vera A. E., Alvin J. Chan, Peter G. Van Sluis, Huib N. Caron, Carel J. M. Van Noesel, and Rogier Versteeg. 2003. “High Delta-like 1 Expression in a Subset of Neuroblastoma Cell Lines Corresponds to a Differentiated Chromaffin Cell Type.” *International Journal of Cancer. Journal International Du Cancer* 105 (1): 61–69. https://doi.org/10.1002/ijc.11047.

Van Loo, Peter, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, et al. 2010. “Allele-Specific Copy Number Analysis of Tumors.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (39): 16910–15. https://doi.org/10.1073/pnas.1009843107.

Vega-Benedetti, Ana F., Cinthia Saucedo, Patrizia Zavattari, Roberta Vanni, José L. Zugaza, and Luis Antonio Parada. 2017. “PLAGL1: An Important Player in Diverse Pathological Processes.” *Journal of Applied Genetics* 58 (1): 71–78. https://doi.org/10.1007/s13353-016-0355-4.

Velasco, Silvia, Mahmoud M. Ibrahim, Akshay Kakumanu, Görkem Garipler, Begüm Aydin, Mohamed Ahmed Al-Sayegh, Antje Hirsekorn, et al. 2017. “A Multi-Step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells.” *Cell Stem Cell* 20 (2): 205–17.e8. https://doi.org/10.1016/j.stem.2016.11.006.

Verhaak, Roel G. W., Vineet Bafna, and Paul S. Mischel. 2019. “Extrachromosomal Oncogene Amplification in Tumour Pathogenesis and Evolution.” *Nature Reviews. Cancer* 19 (5): 283–88. https://doi.org/10.1038/s41568-019-0128-6.

Veyrieras, Jean-Baptiste, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T. Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K. Pritchard. 2008. “High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.” *PLoS Genetics* 4 (10): e1000214. https://doi.org/10.1371/journal.pgen.1000214.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. “ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers.” *Nature* 457 (7231): 854–58. https://doi.org/10.1038/nature07730.

Voelkerding, Karl V., Shale A. Dames, and Jacob D. Durtschi. 2009. “Next-Generation Sequencing: From Basic Research to Diagnostics.” *Clinical Chemistry* 55 (4): 641–58. https://doi.org/10.1373/clinchem.2008.112789.

Vo, Kieuhoa T., Katherine K. Matthay, John Neuhaus, Wendy B. London, Barbara Hero, Peter F. Ambros, Akira Nakagawara, et al. 2014. “Clinical, Biologic, and Prognostic Differences on the Basis of Primary Tumor Site in Neuroblastoma: A Report From the International Neuroblastoma Risk Group Project.” *Journal of Clinical Oncology*. https://doi.org/10.1200/jco.2014.56.1621.

Wada, M., R. C. Seeger, H. Mizoguchi, and H. P. Koeffler. 1995. “Maintenance of Normal Imprinting of H19 and IGF2 Genes in Neuroblastoma.” *Cancer Research* 55 (15): 3386–88. https://www.ncbi.nlm.nih.gov/pubmed/7614476.

Wagner, James R., Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, and Mathieu Blanchette. 2014. “The Relationship between DNA Methylation, Genetic and Expression Inter-Individual Variation in Untransformed Human Fibroblasts.” *Genome Biology* 15 (2): R37. https://doi.org/10.1186/gb-2014-15-2-r37.

Wang, Kai, Sharon J. Diskin, Haitao Zhang, Edward F. Attiyeh, Cynthia Winter, Cuiping Hou, Robert W. Schnepp, et al. 2011. “Integrative Genomics Identifies LMO1 as a Neuroblastoma Oncogene.” *Nature* 469 (7329): 216–20. https://doi.org/10.1038/nature09609.

Wang, Qin, David I. R. Holmes, Sue M. Powell, Qi L. Lu, and Jonathan Waxman. 2002. “Analysis of Stromal-Epithelial Interactions in Prostate Cancer Identifies PTPCAAX2 as a Potential Oncogene.” *Cancer Letters* 175 (1): 63–69. https://doi.org/10.1016/s0304-3835(01)00703-0.

Wang, Qun, Sharon Diskin, Eric Rappaport, Edward Attiyeh, Yael Mosse, Daniel Shue, Eric Seiser, et al. 2006. “Integrative Genomics Identifies Distinct Molecular Classes of Neuroblastoma and Shows That Multiple Genes Are Targeted by Regional Alterations in DNA Copy Number.” *Cancer Research* 66 (12): 6050–62. https://doi.org/10.1158/0008-5472.CAN-05-4618.

Wang, Xiaojie, Waleed M. Ghareeb, Xingrong Lu, Ying Huang, Shenghui Huang, and Pan Chi. 2019. “Coexpression Network Analysis Linked H2AFJ to Chemoradiation Resistance in Colorectal Cancer.” *Journal of Cellular Biochemistry* 120 (6): 10351–62. https://doi.org/10.1002/jcb.28319.

Wang, Xinchen, and David B. Goldstein. 2020. “Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease.” *American Journal of Human Genetics* 106 (2): 215–33. https://doi.org/10.1016/j.ajhg.2020.01.012.

Wang, Zhibin, Chongzhi Zang, Jeffrey A. Rosenfeld, Dustin E. Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, et al. 2008. “Combinatorial Patterns of Histone Acetylations and Methylations in the Human Genome.” *Nature Genetics* 40 (7): 897–903. https://doi.org/10.1038/ng.154.

Waszak, Sebastian M., Olivier Delaneau, Andreas R. Gschwind, Helena Kilpinen, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, et al. 2015. “Population Variation and Genetic Control of Modular Chromatin Architecture in Humans.” *Cell* 162 (5): 1039–50. https://doi.org/10.1016/j.cell.2015.08.001.

Weischenfeldt, Joachim, Taronish Dubash, Alexandros P. Drainas, Balca R. Mardin, Yuanyuan Chen, Adrian M. Stütz, Sebastian M. Waszak, et al. 2017. “Pan-Cancer Analysis of Somatic Copy-Number Alterations Implicates IRS4 and IGF2 in Enhancer Hijacking.” *Nature Genetics* 49 (1): 65–74. https://doi.org/10.1038/ng.3722.

West, Lisandra E., and Or Gozani. 2011. “Regulation of p53 Function by Lysine Methylation.” *Epigenomics* 3 (3): 361–69. https://doi.org/10.2217/EPI.11.21.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.” *Cell* 153 (2): 307–19. https://doi.org/10.1016/j.cell.2013.03.035.

Wilhelm, Mathias, Hannes Hahne, Mikhail Savitski, Harald Marx, Simone Lemeer, Marcus Bantscheff, and Bernhard Kuster. 2017. “Wilhelm et Al. Reply.” *Nature* 547 (7664): E23. https://doi.org/10.1038/nature22294.

Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, et al. 2014. “Mass-Spectrometry-Based Draft of the Human Proteome.” *Nature* 509 (7502): 582–87. https://doi.org/10.1038/nature13319.

Will, Anja J., Giulia Cova, Marco Osterwalder, Wing-Lee Chan, Lars Wittler, Norbert Brieske, Verena Heinrich, et al. 2017. “Composition and Dosage of a Multipartite Enhancer Cluster Control Developmental Expression of Ihh (Indian Hedgehog).” *Nature Genetics* 49 (10): 1539–45. https://doi.org/10.1038/ng.3939.

Williamson, Christine M., Simon T. Ball, Claire Dawson, Stuti Mehta, Colin V. Beechey, Martin Fray, Lydia Teboul, T. Neil Dear, Gavin Kelsey, and Jo Peters. 2011. “Uncoupling Antisense-Mediated Silencing and DNA Methylation in the Imprinted Gnas Cluster.” *PLoS Genetics* 7 (3): e1001347. https://doi.org/10.1371/journal.pgen.1001347.

Williamson, Christine M., Simon T. Ball, Wade T. Nottingham, Judith A. Skinner, Antonius Plagge, Martin D. Turner, Nicola Powles, et al. 2004. “A Cis-Acting Control Region Is Required Exclusively for the Tissue-Specific Imprinting of Gnas.” *Nature Genetics* 36 (8): 894–99. https://doi.org/10.1038/ng1398.

Wilson, Peter C. G., Max J. Coppes, Hassan Solh, Helen S. L. Chan, Derek Jenkin, Mark L. Greenberg, and Sheila Weitzman. 1991. “Neuroblastoma Stage IV-S: A Heterogeneous Disease.” *Medical and Pediatric Oncology* 19 (6): 467–72. https://onlinelibrary.wiley.com/doi/abs/10.1002/mpo.2950190604.

Wimmer, K., X. X. Zhu, B. J. Lamb, R. Kuick, P. F. Ambros, H. Kovar, D. Thoraval, S. Motyka, J. R. Alberts, and S. M. Hanash. 1999. “Co-Amplification of a Novel Gene, NAG, with the N-Myc Gene in Neuroblastoma.” *Oncogene* 18 (1): 233–38. https://doi.org/10.1038/sj.onc.1202287.

Wit, N. J. W. de, J. Rijntjes, J. H. S. Diepstra, T. H. van Kuppevelt, U. H. Weidle, D. J. Ruiter, and G. N. P. van Muijen. 2005. “Analysis of Differential Gene Expression in Human Melanocytic Tumour Lesions by Custom Made Oligonucleotide Arrays.” *British Journal*

of Cancer 92 (12): 2249–61. https://doi.org/10.1038/sj.bjc.6602612.
Wittkopp, Patricia J., Belinda K. Haerum, and Andrew G. Clark. 2004. “Evolutionary Changes in Cis and Trans Gene Regulation.” *Nature* 430 (6995): 85–88. https://doi.org/10.1038/nature02698.
Witz, Isaac P., and Orlev Levy-Nissenbaum. 2006. “The Tumor Microenvironment in the Post-PAGET Era.” *Cancer Letters* 242 (1): 1–10. https://doi.org/10.1016/j.canlet.2005.12.005.
Wong, A. J., J. M. Ruppert, J. Eggleston, S. R. Hamilton, S. B. Baylin, and B. Vogelstein. 1986. “Gene Amplification of c-Myc and N-Myc in Small Cell Carcinoma of the Lung.” *Science* 233 (4762): 461–64. https://doi.org/10.1126/science.3014659.
Wray, Gregory A. 2007. “The Evolutionary Significance of Cis-Regulatory Mutations.” *Nature Reviews. Genetics* 8 (3): 206–16. https://doi.org/10.1038/nrg2063.
Wright, Fred A., Patrick F. Sullivan, Andrew I. Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, et al. 2014. “Heritability and Genomics of Gene Expression in Peripheral Blood.” *Nature Genetics* 46 (5): 430–37. https://doi.org/10.1038/ng.2951.
Wright, W. E., M. A. Piatyszek, W. E. Rainey, W. Byrd, and J. W. Shay. 1996. “Telomerase Activity in Human Germline and Embryonic Tissues and Cells.” *Developmental Genetics* 18 (2): 173–79. https://doi.org/10.1002/(SICI)1520-6408(1996)18:2%3C173::AID-DVG10%3E3.0.CO;2-3.
Wu, Gang, Alberto Broniscer, Troy A. McEachron, Charles Lu, Barbara S. Paugh, Jared Becksfort, Chunxu Qu, et al. 2012. “Somatic Histone H3 Alterations in Pediatric Diffuse Intrinsic Pontine Gliomas and Non-Brainstem Glioblastomas.” *Nature Genetics* 44 (3): 251–53. https://doi.org/10.1038/ng.1102.
Wu, Sihan, Kristen M. Turner, Nam Nguyen, Ramya Raviram, Marcella Erb, Jennifer Santini, Jens Luebeck, et al. 2019. “Circular ecDNA Promotes Accessible Chromatin and High Oncogene Expression.” *Nature* 575 (7784): 699–703. https://doi.org/10.1038/s41586-019-1763-5.
Wu, Thomas D., and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics* 26 (7): 873–81. https://doi.org/10.1093/bioinformatics/btq057.
Wuyts, W., W. Van Hul, K. De Boulle, J. Hendrickx, E. Bakker, F. Vanhoenacker, F. Mollica, et al. 1998. “Mutations in the EXT1 and EXT2 Genes in Hereditary Multiple Exostoses.” *American Journal of Human Genetics* 62 (2): 346–54. https://doi.org/10.1086/301726.
Yamagishi, H., T. Tsuda, S. Fujimoto, M. Toda, K. Kato, Y. Maekawa, M. Umeno, and M. Anai. 1983. “Purification of Small Polydisperse Circular DNA of Eukaryotic Cells by Use of ATP-Dependent Deoxyribonuclease.” *Gene* 26 (2-3): 317–21. https://doi.org/10.1016/0378-1119(83)90205-6.
Yamazaki, Hiromi, Mikiko Suzuki, Akihito Otsuki, Ritsuko Shimizu, Emery H. Bresnick, James Douglas Engel, and Masayuki Yamamoto. 2014. “A Remote GATA2 Hematopoietic Enhancer Drives Leukemogenesis in inv(3)(q21;q26) by Activating EVI1 Expression.” *Cancer Cell* 25 (4): 415–27. https://doi.org/10.1016/j.ccr.2014.02.008.
Yáñez, Yania, Elena Grau, Virginia C. Rodríguez-Cortez, David Hervás, Enrique Vidal, Rosa Noguera, Miguel Hernández, et al. 2015. “Two Independent Epigenetic Biomarkers Predict Survival in Neuroblastoma.” *Clinical Epigenetics* 7 (February): 16. https://doi.org/10.1186/s13148-015-0054-8.
Yan, Hai, Weishi Yuan, Victor E. Velculescu, Bert Vogelstein, and Kenneth W. Kinzler. 2002. “Allelic Variation in Human Gene Expression.” *Science* 297 (5584): 1143. https://doi.org/10.1126/science.1072545.
Yan, Liang, Qi Li, Juan Yang, and Baoping Qiao. 2018. “TPX2-p53-GLIPR1 Regulatory Circuitry in Cell Proliferation, Invasion, and Tumor Growth of Bladder Cancer.” *Journal of*

Cellular Biochemistry 119 (2): 1791–1803. https://doi.org/10.1002/jcb.26340.
Yao, Douglas W., Luke J. O'Connor, Alkes L. Price, and Alexander Gusev. 2020. "Quantifying Genetic Effects on Disease Mediated by Assayed Gene Expression Levels." *Nature Genetics* 52 (6): 626–33. https://doi.org/10.1038/s41588-020-0625-2.
Yao, Jun, Stanislawa Weremowicz, Bin Feng, Robert C. Gentleman, Jeffrey R. Marks, Rebecca Gelman, Cameron Brennan, and Kornelia Polyak. 2006. "Combined cDNA Array Comparative Genomic Hybridization and Serial Analysis of Gene Expression Analysis of Breast Tumor Progression." *Cancer Research* 66 (8): 4065–78. https://doi.org/10.1158/0008-5472.CAN-05-4083.
Yates, Andrew D., Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, et al. 2020. "Ensembl 2020." *Nucleic Acids Research* 48 (D1): D682–88. https://doi.org/10.1093/nar/gkz966.
Yates, Lucy R., Moritz Gerstung, Stian Knappskog, Christine Desmedt, Gunes Gundem, Peter Van Loo, Turid Aas, et al. 2015. "Subclonal Diversification of Primary Breast Cancer Revealed by Multiregion Sequencing." *Nature Medicine* 21 (7): 751–59. https://doi.org/10.1038/nm.3886.
Ye, B. H., S. Chaganti, C. C. Chang, H. Niu, P. Corradini, R. S. Chaganti, and R. Dalla-Favera. 1995. "Chromosomal Translocations Cause Deregulated BCL6 Expression by Promoter Substitution in B Cell Lymphoma." *The EMBO Journal* 14 (24): 6209–17. https://www.ncbi.nlm.nih.gov/pubmed/8557040.
Yi, Kijong, and Young Seok Ju. 2018. "Patterns and Mechanisms of Structural Variations in Human Cancer." *Experimental & Molecular Medicine* 50 (8): 98. https://doi.org/10.1038/s12276-018-0112-3.
Yin, D., D. Xie, S. Sakajiri, C. W. Miller, H. Zhu, M. L. Popoviciu, J. W. Said, K. L. Black, and H. P. Koeffler. 2006. "DLK1: Increased Expression in Gliomas and Associated with Oncogenic Activities." *Oncogene* 25 (13): 1852–61. https://doi.org/10.1038/sj.onc.1209219.
Young, Richard A. 2011. "Control of the Embryonic Stem Cell State." *Cell* 144 (6): 940–54. https://doi.org/10.1016/j.cell.2011.01.032.
Yuan, Shuai, and Zhaohui Qin. 2012. "Read-Mapping Using Personalized Diploid Reference Genome for RNA Sequencing Data Reduced Bias for Detecting Allele-Specific Expression." *IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE International Conference on Bioinformatics and Biomedicine* 2012 (October): 718–24. https://doi.org/10.1109/BIBMW.2012.6470225.
Yu, Jianming, Gael Pressoir, William H. Briggs, Irie Vroh Bi, Masanori Yamasaki, John F. Doebley, Michael D. McMullen, et al. 2006. "A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness." *Nature Genetics* 38 (2): 203–8. https://doi.org/10.1038/ng1702.
Zack, Travis I., Stephen E. Schumacher, Scott L. Carter, Andre D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, et al. 2013. "Pan-Cancer Patterns of Somatic Copy Number Alteration." *Nature Genetics* 45 (10): 1134–40. https://doi.org/10.1038/ng.2760.
Zaret, Kenneth S., and Jason S. Carroll. 2011. "Pioneer Transcription Factors: Establishing Competence for Gene Expression." *Genes & Development* 25 (21): 2227–41. https://doi.org/10.1101/gad.176826.111.
Zech, L., U. Haglund, K. Nilsson, and G. Klein. 1976. "Characteristic Chromosomal Abnormalities in Biopsies and Lymphoid-Cell Lines from Patients with Burkitt and Non-Burkitt Lymphomas." *International Journal of Cancer. Journal International Du Cancer* 17 (1): 47–56. https://doi.org/10.1002/ijc.2910170108.
Zeid, Rhamy, Matthew A. Lawlor, Evon Poon, Jaime M. Reyes, Mariateresa Fulciniti, Michael A. Lopez, Thomas G. Scott, et al. 2018. "Enhancer Invasion Shapes MYCN-Dependent

Transcriptional Amplification in Neuroblastoma." *Nature Genetics* 50 (4): 515–23. https://doi.org/10.1038/s41588-018-0044-9.
Zhang, Tongwu, Jiyeon Choi, Michael A. Kovacs, Jianxin Shi, Mai Xu, NISC Comparative Sequencing Program, Melanoma Meta-Analysis Consortium, et al. 2018. "Cell-Type-Specific eQTL of Primary Melanocytes Facilitates Identification of Melanoma Susceptibility Genes." *Genome Research* 28 (11): 1621–35. https://doi.org/10.1101/gr.233304.117.
Zhang, Xiaoyang, Peter S. Choi, Joshua M. Francis, Marcin Imielinski, Hideo Watanabe, Andrew D. Cherniack, and Matthew Meyerson. 2016. "Identification of Focally Amplified Lineage-Specific Super-Enhancers in Human Epithelial Cancers." *Nature Genetics* 48 (2): 176–82. https://doi.org/10.1038/ng.3470.
Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, et al. 2010. "Mixed Linear Model Approach Adapted for Genome-Wide Association Studies." *Nature Genetics* 42 (4): 355–60. https://doi.org/10.1038/ng.546.
Zhou, Jian, and Olga G. Troyanskaya. 2015. "Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model." *Nature Methods* 12 (10): 931–34. https://doi.org/10.1038/nmeth.3547.
Zhou, Jun, Bernardo Lemos, Erik B. Dopman, and Daniel L. Hartl. 2011. "Copy-Number Variation: The Balance between Gene Dosage and Expression in Drosophila Melanogaster." *Genome Biology and Evolution* 3: 1014–24. https://doi.org/10.1093/gbe/evr023.
Zhu, Anqi, Joseph G. Ibrahim, and Michael I. Love. 2019. "Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences." *Bioinformatics* 35 (12): 2084–92. https://doi.org/10.1093/bioinformatics/bty895.
Zimmerman, Mark W., Yu Liu, Shuning He, Adam D. Durbin, Brian J. Abraham, John Easton, Ying Shao, et al. 2018. "MYC Drives a Subset of High-Risk Pediatric Neuroblastomas and Is Activated through Mechanisms Including Enhancer Hijacking and Focal Enhancer Amplification." *Cancer Discovery* 8 (3): 320–35. https://doi.org/10.1158/2159-8290.CD-17-0993.

Publications

Koche, R. P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I. C., Maag, J., Chamorro, R., Munoz-Perez, N., Puiggròs, M., Dorado Garcia, H., Bei, Y., Röefzaad, C., Bardinet, V., Szymansky, A., Winkler, A., Thole, T., Timme, N., Kasack, K., Fuchs, S., … Henssen, A. G. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nature Genetics, 52(1), 29–34. https://doi.org/10.1038/s41588-019-0547-z